LETTRES
SORBONNE
UNIVERSITÉ

RAPPORT ANALYSE DE DONNEES

Par THEVERIN Stelline

Année Universitaire : 2025-2026

SOMMAIRE :

SÉANCE 2 : Questions de cours :

1) Quel est le positionnement de la géographie par rapport à la statistique ?

Beaucoup de géographes défendent la position du hasard comme l'origine de toute chose, et tiennent à leur position de "carrefour" entre les sciences et les lettres, un paradoxe quand on en arrive à la statistique, donc. Ce paradoxe est renforcé lorsque la géographie souhaite également se revendiquer en tant que science, sans avoir la rigueur, ou la rigidité scientifique. Dans le cas de la géographie humaine notamment, cette rigidité n'est pas essentielle, voire pas conseillée. Cet aspect de nuance est renforcé par l'absence de cours de mathématiques en première année de licence. Cette position peut se heurter à un mur lorsque les statistiques sont abordées, puisqu'on parle alors de chiffres purs et durs, d'indicateurs, de faits qui sont à analyser. Beaucoup de jeunes géographes appréhendent alors cet aspect qui fait partie intégrante de la géographie (puisqu'il n'est pas d'analyse dans aucune source sûre, et que les analyses statistiques sont fondamentales en ce sens). Le hasard en lui-même est une vision philosophique, c'est à chacun de se positionner, mais le géographe entretient forcément un lien avec les statistiques : il en a besoin, mais dans la mesure ou la géographie française tend à suivre une approche vidalienne et à la voir comme une méthode et pas uniquement une science, il se doit d'avoir une certaine flexibilité propre aux lettres, tout en ayant une rigueur presque scientifique dans l'analyse pour asseoir sa véracité.

2) Le hasard existe-t-il en géographie ?

Le hasard est avant tout un concept philosophique, mais il existe rarement de manière pure et dure dans la géographie. Cette dernière tente d'expliquer les procédés, les évènements, les mécanismes complexes. Dans le cas du développement par exemple, il est question d'étudier les procédés utilisés, les politiques, les cultures et les éléments annexes qui permettent d'arriver à une situation. Il peut y avoir des éléments inattendus sur le moment, des comportements de la part d'individus. Avec la géographie humaine, on ne peut pas prédire avec exactitude une situation, cependant, on peut dégager des tendances et des certitudes globales. L'attitude la plus vraisemblable à avoir. Les lois du hasard n'excluent pas un écart sur cette vraisemblance. C'est là qu'on retrouve un raisonnement géographique inchangé depuis ses prémices : le raisonnement multiscalaire. Ce dernier, appliqué à notre cas de figure, peut vouloir dire qu'à l'échelle d'un état, selon une situation

donnée, certaines options ou certaines tendances sont visibles ou vraisemblables, mais à une échelle locale, des dynamiques inverses peuvent s'observer.

3) Quels sont les types d'information géographique ?

Il y a deux séries statistiques possibles pour les types d'information géographique : il peut y avoir les données territoriales claires et précises d'ensembles délimités par des éléments de géographie humaine, que ce soit la population, les caractéristiques sociales, économiques… ou de géographie physique comme la température ou le volume des précipitations.

D'un autre côté, il peut s'agir d'étudier la morphologie même des ensembles délimités.

4) Quels sont les besoins de la géographie au niveau de l'analyse de données ?

La plupart du temps, le géographe ne produit pas ses propres données d'analyse, mais laisse le soin aux organismes gouvernementaux, officiels, ou aux institutions ce travail. Le géographe va exploiter ces données. Même en géographie physique, quelques mesures sont faites, mais la majorité du travail est laissée aux topographes ou aux géologues entre autres. Le géographe va cependant modifier la nomenclature et les métadonnées, il se doit de les organiser pour les exploiter correctement.

Pour la Nomenclature : avant de faire une analyse, il faut produire la donnée et la nomenclature est un ensemble de définitions préalables au recueil de l'information. Elles peuvent être hiérarchisées ou non, s'additionner, se soustraire entre elles, modulées afin de répondre précisément au besoin du géographe afin qu'il obtienne une information plus ou moins agrégée.

5) Quelles sont les différences entre la stat descriptive et la stat explicative ?

Les deux se distinguent par leur finalité et leur portée analytique. La statistique descriptive vise essentiellement à résumer et organiser l'information brute afin de rendre les données lisibles et compréhensibles, à travers des outils tels que les moyennes, les médianes, les écarts-types ou encore les représentations graphiques. Elle ne cherche pas à établir de relations causales, mais à fournir une photographie fidèle d'un phénomène observé.

La statistique explicative, quant à elle, s'inscrit dans une démarche analytique plus ambitieuse : elle mobilise des modèles et des tests pour identifier, mesurer et interpréter les liens entre variables, dans le but de comprendre les mécanismes sous-jacents et d'expliquer pourquoi un phénomène se produit. Autrement dit, la première décrit « ce qui est », tandis que la seconde cherche à rendre intelligible « pourquoi cela est ».

6) Quels sont les types de visualisation de données en géographie et comment les choisir ?

Les visualisations de données en géographie se déclinent en plusieurs catégories selon la nature des phénomènes étudiés. Les cartes thématiques constituent l'outil central :

- les choroplèthes traduisent des valeurs par des nuances de couleur, les cartes de points localisent des occurrences précises.
- Les cartes de flux représentent des mouvements ou des échanges entre territoires.
- Outre les cartes, les anamorphoses déforment l'espace pour mettre en évidence des disparités.
- Les graphiques statistiques (histogrammes, diagrammes en barres, courbes temporelles) offrent une lecture complémentaire des dynamiques quantitatives.

Ces différents supports permettent de rendre visibles des dimensions spatiales, temporelles ou relationnelles qui seraient difficiles à saisir autrement.

Le choix d'une visualisation dépend avant tout de la question de recherche et du type de données disponibles. Les cartes choroplèthes sont adaptées aux variables continues ou aux taux rapportés à une unité spatiale, tandis que les cartes de flux s'imposent pour représenter des mobilités ou des échanges. Les cartes de points conviennent aux phénomènes ponctuels (localisation d'infrastructures, événements), et les graphiques temporels éclairent les évolutions dans la durée. Ainsi, sélectionner une visualisation suppose de croiser la nature des données (qualitatives, quantitatives, spatiales, temporelles) avec l'objectif analytique (décrire, comparer, expliquer), afin de produire une représentation à la fois lisible, pertinente et scientifiquement rigoureuse.

7) Quelles sont les méthodes d'analyse de données possibles ?

Quand on parle des méthodes d'analyse de données, il s'agit en fait de différentes manières de « faire parler » les chiffres ou les informations qu'on a collecté. Certaines approches sont assez simples, comme les statistiques descriptives qui résument les données avec des moyennes, des pourcentages ou des graphiques. D'autres vont plus loin, comme l'analyse exploratoire qui cherche à repérer des tendances ou des corrélations inattendues, ou encore les méthodes explicatives qui utilisent des modèles pour comprendre les relations entre variables. On peut aussi mobiliser des techniques plus spécialisées, comme l'analyse spatiale en géographie pour voir comment un phénomène se répartit dans l'espace, ou l'analyse temporelle pour suivre son évolution dans le temps. En pratique, le choix dépend surtout de ce qu'on veut savoir : décrire, comparer, expliquer ou prédire.

8) Comment définir: population statistique ? Individu statistique ? Caractère statistique ? Modalités statistiques ? Quels sont les types de caractères ? Y-a-t-il une hiérarchie entre eux ?

Population statistique La population statistique désigne l'ensemble des éléments ou des individus sur lesquels porte une étude. Par exemple, si l'on s'intéresse aux habitudes de lecture des étudiants d'une université, la population statistique correspond à tous les étudiants inscrits dans cette université. Elle constitue le cadre global de l'analyse et permet de définir le champ d'application des résultats.

Individu statistique Un individu statistique est chacun des éléments qui composent la population étudiée. Dans l'exemple précédent, chaque étudiant est un individu statistique. Si l'on change de contexte, pour une enquête sur les communes françaises, chaque commune devient un individu statistique. L'individu est donc l'unité de base sur laquelle on observe des caractéristiques.

Caractère statistique Le caractère statistique est la propriété ou la variable que l'on observe chez les individus. Il peut s'agir de l'âge, du revenu, du nombre de livres lus par mois, ou encore du type de logement occupé. Par exemple, dans une enquête sur les étudiants, le caractère étudié pourrait être « nombre d'heures consacrées à la lecture par semaine ».

Modalités statistiques Les modalités statistiques sont les différentes valeurs ou catégories que peut prendre un caractère. Si le caractère est « type de logement », les modalités peuvent être «

appartement », « maison », « résidence universitaire ». Si le caractère est « nombre de livres lus par mois », les modalités sont les valeurs numériques possibles (0, 1, 2, 3, etc.). Elles permettent de décomposer la diversité des observations.

Types de caractères On distingue principalement deux types de caractères : qualitatifs et quantitatifs. Les caractères qualitatifs décrivent une qualité ou une catégorie (ex. : couleur des yeux, type de logement), tandis que les caractères quantitatifs expriment une mesure numérique (ex. : âge, revenu, nombre d'enfants). Les caractères quantitatifs peuvent eux-mêmes être discrets (valeurs entières comme le nombre d'enfants) ou continus (valeurs mesurables comme la taille ou le revenu).

Hiérarchie entre les caractères Il existe une hiérarchie implicite entre les types de caractères, liée à leur niveau de précision et aux traitements statistiques possibles. Les caractères qualitatifs nominaux (comme la couleur des yeux) permettent surtout des regroupements et des comparaisons de fréquences. Les caractères qualitatifs ordonnés (comme un niveau de satisfaction : faible, moyen, élevé) ajoutent une dimension de classement. Les caractères quantitatifs, enfin, offrent le plus haut degré de précision puisqu'ils permettent des calculs (moyennes, écarts-types, corrélations). Ainsi, plus un caractère est quantitatif et mesurable, plus il ouvre la voie à des analyses fines et complexes.

9) Comment mesurer une amplitude et une densité ?

L'amplitude et la densité sont deux façons complémentaires de caractériser une série statistique. L'amplitude correspond à l'écart entre la valeur la plus grande et la valeur la plus petite observée : par exemple, si l'âge des individus d'un groupe varie entre 18 et 65 ans, l'amplitude est de $65-18=47$ ans. Elle donne une idée de la dispersion globale des données. La densité, quant à elle, est utilisée surtout pour comparer des classes de l' histogramme lorsque les intervalles n'ont pas la même largeur. Elle se calcule en divisant l'effectif d'une classe par la largeur de son intervalle, ce qui permet de représenter correctement la fréquence relative des données. Par exemple, si une classe regroupe 20 individus sur un intervalle de 10 unités, sa densité est $20/10=2$. Ainsi, l'amplitude mesure l'étendue totale des valeurs, tandis que la densité permet de comparer la concentration des observations entre classes de tailles différentes.

10) A quoi servent les formules de Sturges et Yule ?

Les formules de Sturges et de Yule ont pour objectif d'aider le statisticien à déterminer un nombre de classes pertinent lors de la construction d'un histogramme ou d'un tableau de fréquences, afin de représenter une distribution de données de manière lisible et équilibrée. La règle de Sturges, élaborée en 1926, propose de fixer le nombre de classes en fonction de la taille de l'échantillon selon la formule $k=1+\log_2(n)$, où n est le nombre d'observations. Ainsi, pour un échantillon de 100 individus, on obtient $k=1+\log_2(100)\approx 8$ classes. Cette approche évite un découpage arbitraire et garantit une représentation synthétique. La formule de Yule, quant à elle, affine ce principe en tenant compte de l'étendue des données (amplitude) et de la taille de l'échantillon, ce qui permet d'ajuster le nombre de classes lorsque la distribution est particulièrement large ou hétérogène. Par exemple, pour une série de 200 observations dont les valeurs s'étendent de 0 à 100, Yule recommande davantage de classes qu'une application stricte de Sturges, afin de ne pas masquer les variations internes. Ces formules servent donc à rationaliser le choix du nombre de classes, en conciliant précision statistique et clarté graphique, et constituent des outils fondamentaux pour la statistique descriptive.

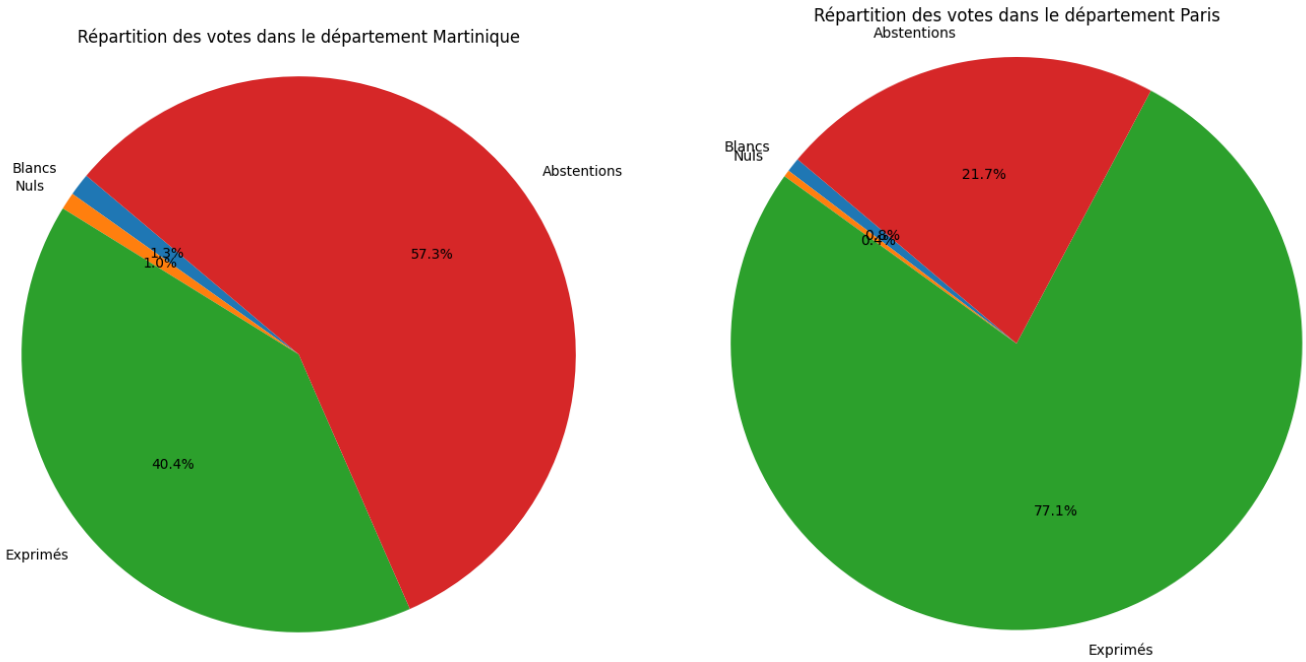
11) Comment définir un effectif ? Comment calculer une fréquence et une fréquence cumulée ? Qu'est-ce qu'une distribution statistique ?

Effectif: En statistique, l'effectif désigne le nombre d'individus ou d'observations correspondant à une modalité donnée d'un caractère. Par exemple, si l'on étudie le caractère « type de logement » et que 30 étudiants vivent en résidence universitaire, l'effectif de cette modalité est 30. L'effectif total correspond à la taille de la population étudiée.

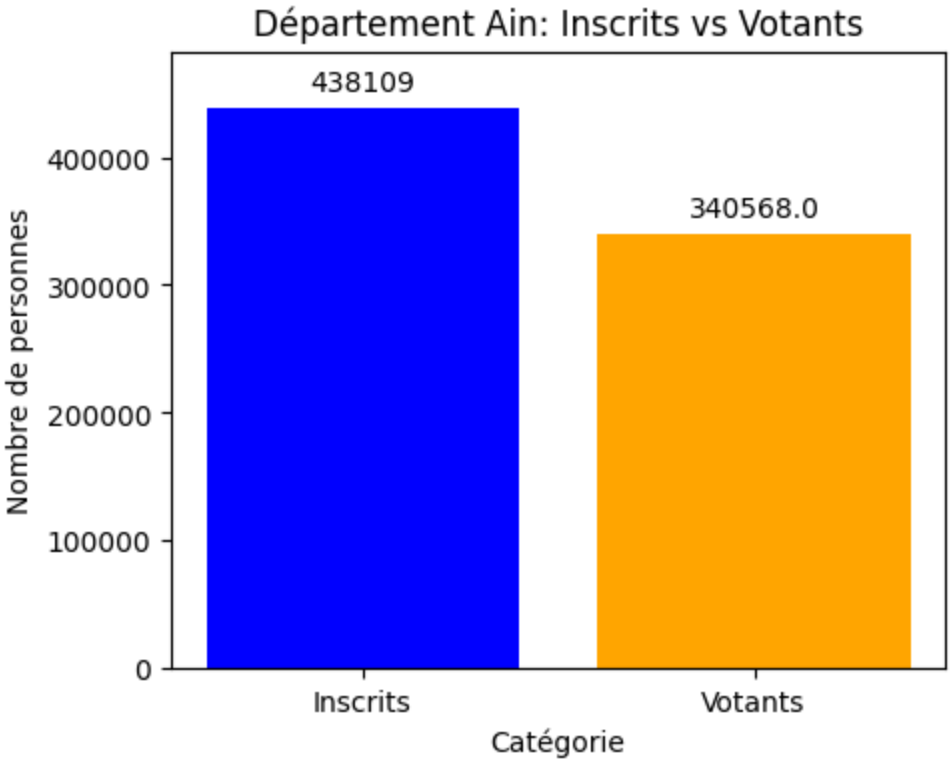
Fréquence et fréquence cumulée La fréquence exprime la proportion d'individus appartenant à une modalité par rapport à l'ensemble de la population. Elle se calcule en divisant l'effectif de la modalité par l'effectif total, puis en multipliant éventuellement par 100 pour obtenir un pourcentage. Par exemple, si 30 étudiants sur 120 vivent en résidence universitaire, la fréquence est $30/120=0{,}25$, soit 25 %. La fréquence cumulée, quant à elle, additionne les fréquences des modalités successives (souvent ordonnées) afin de montrer la progression de la répartition. Ainsi, si les fréquences des trois premières modalités sont 25 %, 40 % et 20 %, la fréquence cumulée après la troisième modalité est $25+40+20=85\%$.

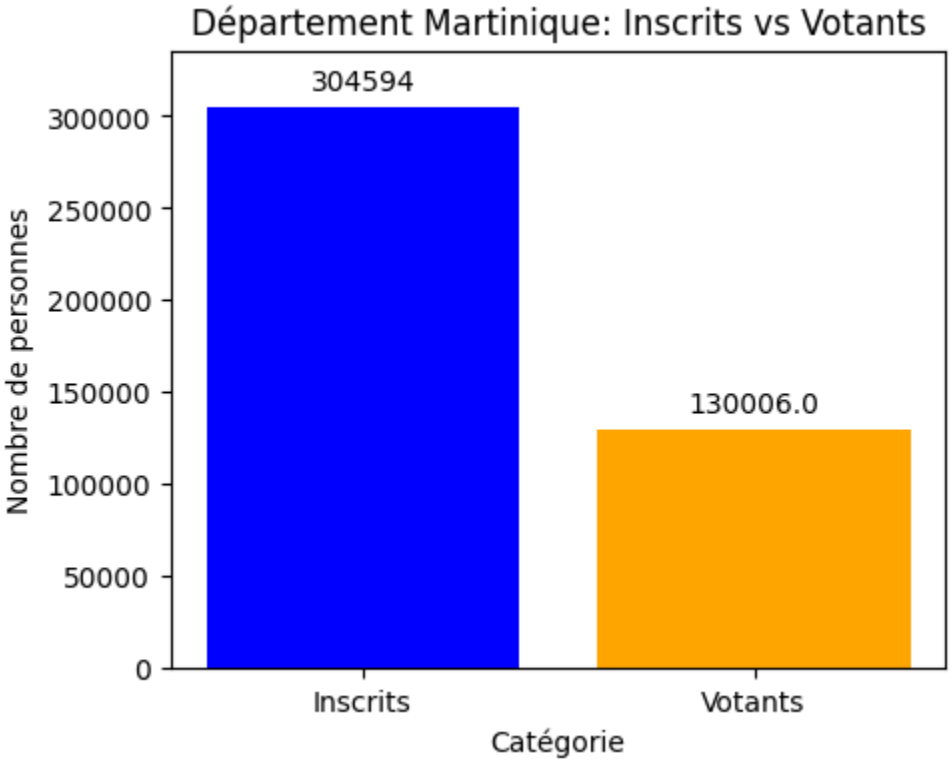
Distribution statistique Une distribution statistique est la manière dont les effectifs ou fréquences se répartissent entre les différentes modalités d'un caractère. Elle peut être présentée sous forme de tableau ou de graphique (histogramme, diagramme en barres, polygone de fréquences). Par exemple, la distribution des types de logement des étudiants montre comment les effectifs se répartissent entre « appartement », « maison » et « résidence universitaire ». La distribution statistique est donc un outil essentiel pour visualiser et analyser la structure d'un phénomène.

Application Python :



Les diagrammes en secteurs (camemberts) fournis synthétisent la répartition relative des catégories de vote et doivent être interprétés prioritairement en pourcentages plutôt qu'en valeurs absolues : un secteur large indique une part importante du total du département, mais sa signification pratique dépend de la taille de l'électorat. Pour la Martinique, dont l'échantillon électoral est généralement de faible effectif, les parts exprimées par le camembert sont particulièrement sensibles aux variations aléatoires — une petite différence en nombre de voix peut se traduire par une grande variation en pourcentage ; il convient donc de mentionner les effectifs bruts et, si possible, les intervalles de confiance des proportions pour montrer l'incertitude associée. À l'inverse, le camembert de Paris, qui reflète un électorat beaucoup plus vaste, peut afficher des différences de pourcentage modestes mais correspondant à des milliers de voix : ces écarts ont une portée pratique plus importante et doivent être évalués à la fois en relatif et en absolu.


Département Ain: Inscrits vs Votants
438109
340568.0
400000
300000
200000
100000
0
Nombre de personnes
Inscrits
Votants
Catégorie


Département Martinique: Inscrits vs Votants
304594
130006.0
300000
250000
200000
150000
100000
50000
0
Nombre de personnes
Inscrits
Votants
Catégorie

Les histogrammes générés pour chaque département constituent une série d'agrégats visuels dont le nombre correspond au nombre de départements distincts présents dans le jeu de données : pour chaque département on produit un graphique séparé, de sorte que le nombre total d'histogrammes est égal au nombre de lignes (ou d'identifiants départementaux) uniques du tableau. Chaque histogramme est composé de deux modalités essentielles — les barres « Inscrits » et « Votants » — qui représentent respectivement l'effectif des personnes inscrites et celui des votants sur la même échelle absolue ; l'axe vertical renseigne les effectifs et l'axe horizontal les catégories (Inscrits / Votants). L'analyse visuelle de ces histogrammes doit porter sur deux éléments principaux : (i) la hauteur relative des barres, qui donne l'information brute sur l'ampleur de l'électorat et la participation, et (ii) le rapport entre les deux barres (Votants / Inscrits), indicateur de taux de participation utile pour comparer des départements de tailles différentes. Il convient aussi de contrôler la lisibilité (étiquettes, unités), la présence d'outliers ou de valeurs manquantes et, le cas échéant, de normaliser les représentations en pourcentages pour faciliter la comparaison.

Concernant la Martinique et l'Ain, l'observation pertinente est de distinguer la dimension absolue (taille de l'électorat) et la dimension relative (taux de participation) : la Martinique, en tant que collectivité d'outre-mer, se caractérise typiquement par des effectifs absolus plus faibles que la plupart des départements métropolitains ; son histogramme présentera donc des barres d'amplitude moindre — mais cela ne permet pas à lui seul d'inférer une faible mobilisation. L'information décisive est le rapport Votants/Inscrits : si la barre « Votants » représente une part importante des « Inscrits » (rapport élevé), alors la Martinique affiche une participation proportionnellement forte malgré une taille réduite ; si au contraire la part est faible, cela signale une mobilisation moindre. Pour l'Ain, département métropolitain de taille intermédiaire à élever, l'histogramme montrera des effectifs absolus supérieurs ; l'attention se portera sur la stabilité relative du taux de participation et sur d'éventuelles différences structurelles par rapport à la moyenne nationale (écarts prononcés, présence d'anomalies locales). En résumé : l'interprétation correcte exige de combiner l'examen des histogrammes absolus avec le calcul du taux de participation (Votants / Inscrits) et, lorsque l'on compare des départements hétérogènes, de privilégier les représentations en pourcentages pour éviter des conclusions biaisées par la taille des populations.

SÉANCE 3 : Questions de cours :

1) Quel caractère est le plus général : le caractère quantitatif ou le caractère qualitatif ? Justifier pourquoi.

Le caractère qualitatif est considéré comme le plus général, car il englobe toutes les situations où l'on décrit une propriété ou une catégorie sans nécessairement recourir à une mesure numérique. En effet, tout caractère statistique peut être ramené à une distinction qualitative (par exemple, « âge » peut être traduit en classes d'âge : jeune, adulte, senior), tandis que l'inverse n'est pas toujours possible : un caractère purement qualitatif, comme la couleur des yeux ou le type de logement, ne peut pas être transformé en valeur numérique sans perdre sa signification. Le caractère quantitatif, qui repose sur des mesures chiffrées (taille, revenu, nombre d'enfants), constitue donc un cas particulier du caractère qualitatif, puisqu'il suppose une opération de mesure et une échelle numérique. C'est pourquoi, dans une hiérarchie conceptuelle, le caractère qualitatif est plus général : il inclut à la fois les variables catégorielles et, par extension, les variables mesurées qui peuvent être classées ou regroupées.

2) Que sont les caractères quantitatifs discrets et caractères quantitatifs continus ? Pourquoi les distinguer ?

Les caractères quantitatifs se divisent en deux grandes catégories : **discrets** et **continus**. Un caractère quantitatif discret est une variable numérique qui ne peut prendre que des valeurs entières et dénombrables. Par exemple, le nombre d'enfants dans une famille ou le nombre de pièces dans un logement : on ne peut pas avoir 2,5 enfants ou 3,7 pièces, les valeurs possibles sont limitées à des entiers. À l'inverse, un caractère quantitatif continu peut prendre toutes les valeurs possibles dans un intervalle donné, y compris des fractions ou des décimales. La taille d'une personne, le revenu mensuel ou la température sont des exemples de variables continues, car elles peuvent varier de manière infiniment fine.

Il est important de les distinguer car les méthodes d'analyse et de représentation diffèrent selon la nature du caractère. Les caractères discrets se prêtent bien aux tableaux de fréquences et aux diagrammes en barres, tandis que les caractères continus nécessitent souvent un regroupement en

classes pour être représentés, par exemple dans un histogramme. Cette distinction conditionne donc le choix des outils statistiques et graphiques, et garantit une interprétation correcte des données.

3) Paramètres de position

a. Pourquoi existe-t-il plusieurs types de moyenne ?

Dans les paramètres de position, il existe plusieurs types de moyenne car chacune répond à une logique spécifique et permet de mieux représenter la tendance centrale selon la nature des données et leur distribution. La **moyenne arithmétique** est la plus courante : elle additionne toutes les valeurs et les divise par le nombre d'observations, ce qui donne une mesure globale mais sensible aux valeurs extrêmes. La **moyenne géométrique**, qui consiste à prendre la racine n-ième du produit des valeurs, est plus adaptée aux phénomènes de croissance ou de proportion (par exemple, des taux d'évolution annuels). La **moyenne harmonique**, calculée comme l'inverse de la moyenne des inverses, est pertinente lorsqu'on travaille avec des vitesses ou des ratios. Enfin, la **moyenne quadratique** (ou RMS) met davantage de poids sur les valeurs élevées et s'utilise notamment en physique ou en ingénierie. Ainsi, la coexistence de plusieurs types de moyenne reflète la diversité des contextes d'analyse : selon que l'on cherche à neutraliser l'effet des extrêmes, à représenter des évolutions multiplicatives ou à comparer des rapports, le choix de la moyenne change pour garantir une mesure représentative et scientifiquement rigoureuse.

b. Pourquoi calculer une médiane ?

La médiane est calculée pour identifier la valeur centrale d'une série statistique et ainsi fournir une mesure de tendance qui n'est pas influencée par les valeurs extrêmes. Contrairement à la moyenne arithmétique, qui peut être fortement déformée par quelques observations très élevées ou très faibles, la médiane divise la population en deux groupes de taille égale : 50 % des individus ont une valeur inférieure ou égale à la médiane, et 50 % une valeur supérieure ou égale. Par exemple, si l'on étudie les revenus mensuels de 11 personnes et que ceux-ci varient de 800 € à 10 000 €, la moyenne sera tirée vers le haut par les revenus les plus élevés, tandis que la médiane donnera une image plus représentative du revenu « typique » de la population. En ce sens, la médiane est

particulièrement utile pour décrire des distributions asymétriques ou hétérogènes, et elle constitue un paramètre de position robuste, garantissant une interprétation plus fidèle de la réalité sociale ou économique.

c. Quand est-il possible de calculer un mode ?

Le mode est une mesure de position qui correspond à la valeur ou à la modalité la plus fréquente dans une distribution statistique. Il est possible de le calculer dès lors que l'on dispose d'un caractère pour lequel les effectifs ou les fréquences sont connus. Concrètement, le mode peut être déterminé aussi bien pour des **caractères qualitatifs** (par exemple, la couleur des yeux la plus répandue dans une classe) que pour des **caractères quantitatifs discrets** (par exemple, le nombre d'enfants le plus courant dans une population). En revanche, pour des **caractères continus**, le mode n'est pas directement calculable sur les valeurs brutes : il nécessite un regroupement en classes et se définit alors comme la classe modale, c'est-à-dire celle qui possède l'effectif le plus élevé. Ainsi, le mode est possible à calculer dès qu'une distribution statistique présente une concentration notable autour d'une valeur ou d'une catégorie, ce qui en fait un indicateur simple mais utile pour identifier la tendance dominante d'un phénomène.

4) Paramètres de concentration

a. Quel est l'intérêt de la médiale et de l'indice de C. Gini ?

Dans les paramètres de concentration, la **médiane** et l'**indice de Gini** jouent des rôles complémentaires pour analyser la répartition d'un caractère statistique, en particulier lorsqu'il s'agit de revenus, de patrimoines ou de toute variable économique. La médiane est un indicateur de position robuste : elle partage la population en deux groupes égaux et permet d'identifier le niveau « typique » sans être influencée par les valeurs extrêmes. Par exemple, dans une étude sur les salaires, la médiane indique le revenu au-dessous duquel se situent 50 % des salariés, ce qui donne une vision plus représentative que la moyenne dans des distributions très inégalitaires. L'indice de Gini, quant à lui, est un paramètre de concentration qui mesure le degré d'inégalité dans la distribution : il varie entre 0 (égalité parfaite, où tous les individus ont la même valeur) et 1 (inégalité maximale, où une seule personne concentre toute la valeur). Ainsi, si deux populations ont la même médiane de revenu, l'indice de Gini permet de distinguer celle où les écarts internes

sont plus marqués. L'intérêt de ces deux outils est donc de fournir une lecture nuancée : la médiane décrit le niveau central, tandis que l'indice de Gini évalue la dispersion et la concentration, offrant ensemble une compréhension plus complète des inégalités.

5) Paramètres de dispersion

a. Pourquoi calculer une variance à la place de l'écart à la moyenne ? Pourquoi la remplacer par l'écart type ?

La question des paramètres de dispersion en statistique met en lumière la nécessité de disposer d'indicateurs robustes et comparables pour mesurer la variabilité d'un ensemble de données. L'« écart à la moyenne » — c'est-à-dire la différence entre chaque valeur et la moyenne arithmétique — constitue une première approche intuitive. Cependant, si l'on se contente de sommer ces écarts, le résultat est nul, puisque les valeurs positives et négatives se compensent. Pour contourner cette difficulté, il est nécessaire de transformer ces écarts afin d'obtenir une mesure non nulle et représentative de la dispersion globale.

La variance répond précisément à ce besoin. En élevant chaque écart au carré, on supprime les signes négatifs et on accorde un poids plus important aux écarts les plus grands. La variance devient ainsi une mesure mathématiquement cohérente, qui reflète la dispersion des données autour de la moyenne. Elle possède en outre des propriétés algébriques utiles, notamment dans le cadre des modèles probabilistes et des inférences statistiques, où elle intervient dans la formulation des lois de probabilité et dans le calcul des erreurs.

Cependant, la variance présente un inconvénient majeur : elle est exprimée dans une unité qui est le carré de celle des données initiales. Par exemple, si les données sont en mètres, la variance est en mètres carrés, ce qui rend son interprétation directe peu intuitive. C'est pourquoi l'on préfère souvent utiliser l'écart type, qui correspond à la racine carrée de la variance. L'écart type conserve les qualités mathématiques de la variance tout en ramenant la mesure dans l'unité d'origine des données. Il devient ainsi un indicateur plus lisible et plus facilement mobilisable pour comparer des dispersions ou pour communiquer des résultats.

En somme, la variance constitue un outil théorique indispensable, mais l'écart type en est la traduction pratique et interprétable. L'un et l'autre forment un couple conceptuel : la variance

assure la rigueur mathématique, tandis que l'écart type garantit l'intelligibilité et l'usage opérationnel.

b. Pourquoi calculer l'étendue ?

L'étendue est l'un des indicateurs les plus simples de la dispersion d'une série statistique : elle se définit comme la différence entre la valeur maximale et la valeur minimale observée. Son intérêt réside dans sa capacité à fournir immédiatement une idée de l'amplitude totale des données. En un seul calcul, elle permet de situer l'écart global entre les extrêmes, ce qui peut être utile pour comparer rapidement des distributions ou pour détecter la présence de valeurs atypiques.

Cependant, l'étendue présente des limites importantes. Elle ne prend en compte que deux observations — le minimum et le maximum — et ignore complètement la répartition des valeurs intermédiaires. Ainsi, deux séries peuvent avoir la même étendue mais des dispersions internes très différentes. De plus, l'étendue est particulièrement sensible aux valeurs aberrantes : un seul point extrême peut considérablement augmenter la mesure et donner une impression de variabilité exagérée.

Dans une perspective académique, l'étendue est donc considérée comme un indicateur élémentaire, souvent utilisé en complément d'autres mesures plus robustes comme la variance, l'écart type ou l'écart interquartile. Elle joue un rôle introductif et descriptif, mais ne suffit pas à elle seule pour caractériser la dispersion d'un ensemble de données. En somme, calculer l'étendue permet de saisir rapidement l'amplitude brute d'une distribution, mais son interprétation doit être nuancée et replacée dans un cadre plus large d'analyse statistique.

c. À quoi sert-il de créer un quantile ? Quel(s) est (sont) le(s) quantile(s) le(s) plus utilisé(s) ?

La notion de **quantile** en statistique répond à un besoin fondamental : découper une distribution en intervalles de taille égale afin de mieux comprendre la répartition des données. En pratique, un quantile est une valeur seuil qui divise l'ensemble des observations en groupes contenant chacun une proportion donnée des données. Cette approche permet de dépasser les simples mesures de tendance centrale (comme la moyenne ou la médiane) et d'examiner la structure interne d'une distribution, notamment sa symétrie, ses concentrations et ses extrêmes.

L'intérêt de créer des quantiles réside dans leur capacité à fournir une vision plus nuancée de la dispersion. Par exemple, les quartiles divisent une série en quatre parties égales, ce qui permet de repérer le « cœur » de la distribution (entre le premier et le troisième quartile) et d'identifier les valeurs extrêmes. Les déciles et les percentiles, quant à eux, offrent une granularité plus fine, utile dans les études démographiques, économiques ou médicales, où l'on cherche à situer un individu ou une observation par rapport à une population entière.

Parmi les quantiles, les plus utilisés sont :

- **La médiane (ou 2e quartile)** : elle partage la distribution en deux moitiés égales et constitue un indicateur robuste de tendance centrale.
- **Les quartiles (Q1, Q2, Q3)** : ils permettent de définir l'intervalle interquartile (Q3 – Q1), qui est une mesure de dispersion moins sensible aux valeurs aberrantes que l'étendue.
- **Les percentiles** : très utilisés en sciences sociales, en médecine et en éducation, ils permettent de situer une observation dans une population (par exemple, un élève au 90e percentile se situe parmi les 10 % les meilleurs).

En somme, les quantiles servent à découper et interpréter une distribution de manière proportionnelle, et les quartiles ainsi que les percentiles sont les plus mobilisés, car ils offrent un équilibre entre lisibilité et précision.

d. Pourquoi construire une boîte de dispersion ? Comment l'interpréter ?

La boîte de dispersion, ou boîte à moustaches, est un outil graphique qui permet de représenter de manière synthétique la distribution d'un ensemble de données. Elle repose sur les quartiles et met en évidence à la fois la tendance centrale, la dispersion et la présence éventuelle de valeurs atypiques. Construire une boîte de dispersion sert avant tout à visualiser l'intervalle interquartile, c'est-à-dire la zone comprise entre le premier et le troisième quartile, qui correspond au cœur de la distribution. La médiane, représentée par une ligne à l'intérieur de la boîte, indique la tendance centrale et permet d'apprécier la symétrie ou l'asymétrie des données. Les moustaches, qui s'étendent généralement jusqu'à une distance de 1,5 fois l'écart interquartile au-delà des quartiles, montrent l'étendue des valeurs considérées comme « normales ». Enfin, les points isolés situés au-

delà des moustaches signalent des valeurs atypiques ou aberrantes. L'interprétation d'une boîte de dispersion consiste donc à lire simultanément la position de la médiane, l'amplitude de la boîte et la présence de valeurs extrêmes, ce qui en fait un instrument particulièrement efficace pour comparer plusieurs distributions et détecter rapidement des phénomènes de variabilité ou d'irrégularité.

6) Paramètres de forme

a. Quelle différence faites-vous entre les moments centrés et les moments absolus ? Pourquoi les utiliser ?

Dans le cadre des paramètres de forme, les moments constituent des outils mathématiques permettant de caractériser la distribution d'une variable aléatoire au-delà des simples mesures de tendance centrale et de dispersion. Ils servent à décrire la symétrie, l'aplatissement ou encore la concentration des données autour de la moyenne. On distingue principalement les **moments centrés** et les **moments absolus**, qui répondent chacun à des objectifs spécifiques.

Les **moments centrés** sont calculés à partir des écarts des valeurs par rapport à la moyenne, élevés à une puissance donnée. Le deuxième moment centré correspond à la variance, le troisième à la mesure de l'asymétrie (ou *skewness*), et le quatrième à l'aplatissement (ou *kurtosis*). Leur intérêt réside dans le fait qu'ils permettent de saisir la forme de la distribution en tenant compte de la moyenne comme point de référence. Ils sont particulièrement utiles pour analyser la symétrie ou la dissymétrie d'une série statistique, ainsi que pour comparer des distributions entre elles.

Les **moments absolus**, quant à eux, reposent sur la valeur absolue des écarts à la moyenne, ce qui évite l'annulation des termes négatifs et limite l'effet des puissances sur les signes. Ils sont souvent utilisés pour des mesures plus robustes, moins sensibles aux valeurs extrêmes, et permettent de définir des indicateurs comme la moyenne des écarts absolus. Contrairement aux moments centrés, ils ne visent pas à décrire la forme globale de la distribution mais plutôt à fournir une mesure de dispersion ou de concentration qui reste interprétable même en présence de données atypiques.

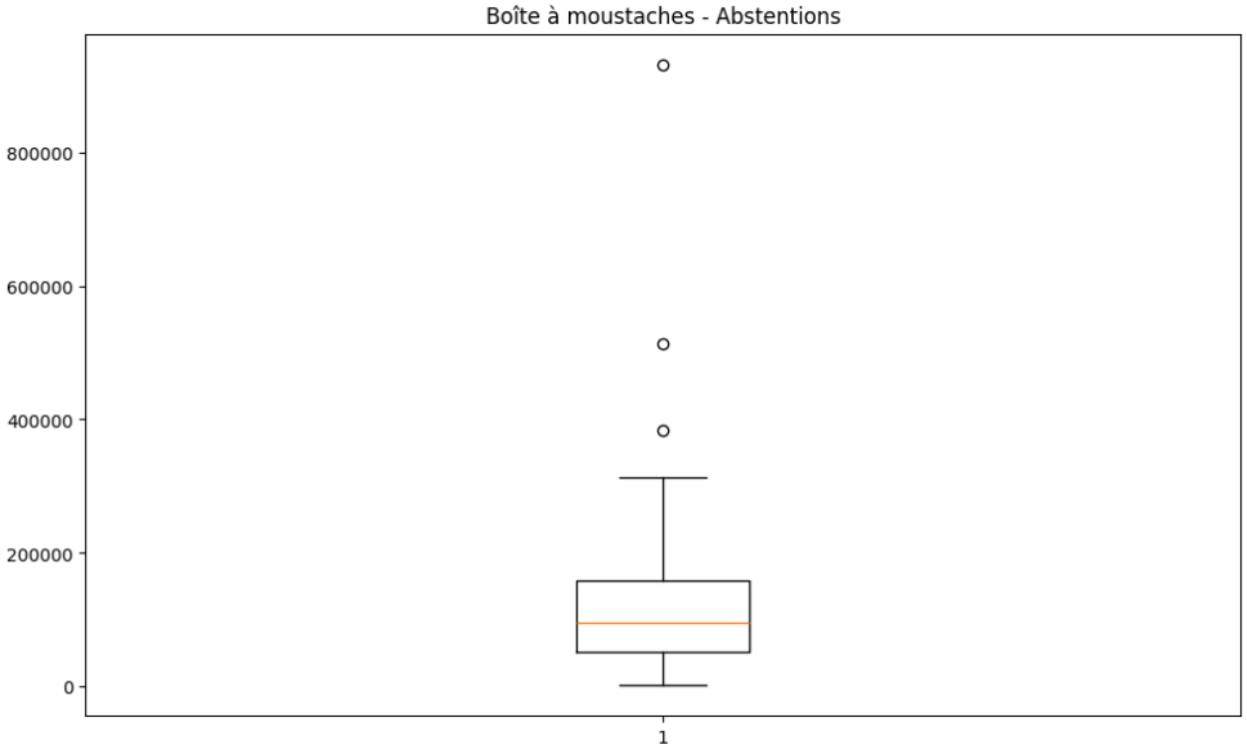
En somme, les moments centrés sont privilégiés lorsqu'il s'agit d'analyser la structure et la forme d'une distribution, notamment son asymétrie et son aplatissement, tandis que les moments absolus offrent une alternative plus robuste pour mesurer la dispersion sans être trop influencés par les

valeurs extrêmes. Leur utilisation conjointe permet d'obtenir une vision plus complète et nuancée des caractéristiques d'une série statistique.

b. Pourquoi vérifier la symétrie d'une distribution et comment faire ?

Vérifier la symétrie d'une distribution est une étape essentielle pour comprendre la structure des données et choisir les outils statistiques les plus adaptés. Une distribution symétrique, comme la loi normale, implique que les valeurs sont réparties de manière équilibrée autour de la moyenne, ce qui facilite l'interprétation et l'utilisation de nombreux tests paramétriques. À l'inverse, une distribution asymétrique peut indiquer la présence de biais, de valeurs extrêmes ou de phénomènes particuliers qu'il convient de prendre en compte. La symétrie influence ainsi directement la pertinence des mesures de tendance centrale : dans une distribution symétrique, la moyenne et la médiane coïncident, tandis que dans une distribution asymétrique, la médiane devient souvent un indicateur plus robuste. Pour vérifier cette symétrie, plusieurs méthodes sont possibles. On peut recourir à une représentation graphique, comme l'histogramme ou la boîte de dispersion, qui permet de visualiser la répartition des données. On peut également utiliser des indicateurs numériques, tels que le coefficient d'asymétrie (ou *skewness*), qui mesure le degré et le sens de la dissymétrie. Enfin, la comparaison entre moyenne et médiane constitue un test simple et intuitif : un écart notable entre ces deux valeurs signale généralement une asymétrie. Ainsi, l'analyse de la symétrie d'une distribution n'est pas seulement descriptive, elle conditionne le choix des méthodes statistiques et la validité des conclusions que l'on peut en tirer.

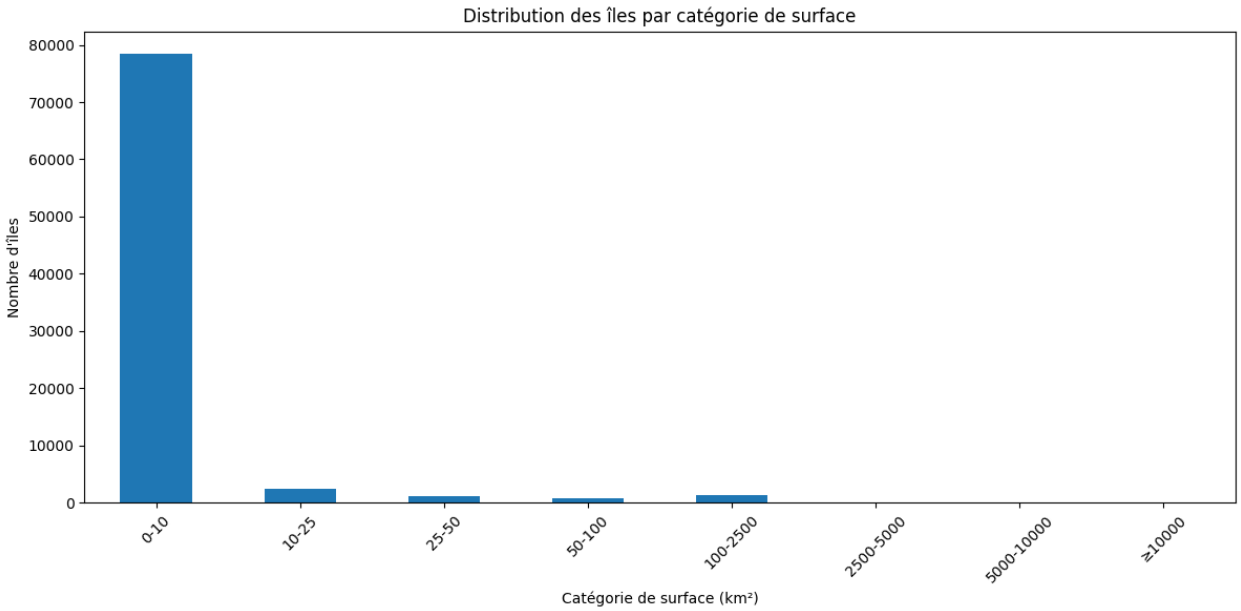
Application Python:



Les boîtes à moustaches fournissent une synthèse visuelle robuste de la distribution des variables (médiane, étendue interquartile, étendue totale et valeurs extrêmes) et permettent d'évaluer rapidement la variabilité et la présence d'anomalies entre départements. Pour les séries « Inscrits » et « Abstentions », on retiendra les points suivants : la position de la médiane renseigne sur le niveau typique d'un département (p. ex. nombre moyen d'inscrits ou part médiane d'abstentions), l'étendue interquartile (IQR) traduit la dispersion centrale (haute IQR = forte hétérogénéité entre départements), et les moustaches / outliers signalent des départements atypiques (petites collectivités ou très grandes circonscriptions).

Inscrits : si la boîte est très étirée et les moustaches longues, cela signifie une forte hétérogénéité des tailles d'électorat (présence de très grands départements comme Paris et de très petits comme certaines collectivités ultramarines). La médiane, comparée à la moyenne, informe sur l'asymétrie (moyenne > médiane indique queue droite, quelques très grands départements tirant la moyenne vers le haut). Les outliers isolent des cas nécessitant un examen (erreurs de saisie ou entités démographiquement particulières).

Abstentions : la boîte et l'IQR traduisent la variabilité des comportements électoraux ; une IQR étroite signale une homogénéité des taux d'abstention entre départements, une IQR large indique des divergences locales. Des outliers élevés en abstention méritent une investigation contextuelle (spécificités locales, événements ponctuels, erreurs de données).



Le diagramme en barres représentant la distribution des îles par classes de surface (extrait de island-index.csv) montre clairement une concentration des effectifs dans les classes de petites surfaces et une longue traîne vers les catégories supérieures : la classe modale se situe parmi les plus petites surfaces (0–100 km² selon le découpage choisi), ce qui indique une distribution fortement asymétrique à droite. Cette configuration suggère que la majorité des îles sont de petite taille tandis que quelques îles très étendues forment une minorité mais influencent fortement les indicateurs de tendance centrale (la moyenne sera nettement supérieure à la médiane). Pour l'interprétation, il faut garder à l'esprit la sensibilité aux bornes de classe : un regroupement différent ou une échelle logarithmique pourrait rendre la structure de la queue plus lisible et éviter la domination visuelle des petites classes. Sur le plan statistique, il est utile de compléter le graphique par des mesures robustes (médiane, IQR) et par un histogramme en densité ou une représentation cumulative pour mieux évaluer la loi sous-jacente (p.ex. log-normale ou loi de puissance). Enfin, avant toute conclusion, vérifier l'absence d'erreurs de saisie ou d'unités incorrectes sur les observations extrêmes, et, si l'objectif est une comparaison relative, normaliser les effectifs par groupe (pourcentages) ou présenter les valeurs en échelle logarithmique afin d'éviter les biais d'interprétation dus à l'hétérogénéité des tailles.

SEANCE 4 : Questions de cours :

1) **Quels critères mettriez-vous en avant pour choisir entre une distribution statistique avec des variables discrètes et une distribution statistique avec des variables continues ?**

Le choix entre une distribution statistique de variables discrètes et une distribution de variables continues dépend avant tout de la nature des données observées. Une variable discrète se caractérise par un nombre fini ou dénombrable de modalités possibles, comme le nombre d'enfants dans une famille ou le résultat d'un lancer de dé. Dans ce cas, la distribution statistique doit refléter cette discontinuité et permettre de représenter la fréquence de chaque modalité distincte. À l'inverse, une variable continue peut prendre une infinité de valeurs dans un intervalle donné, comme la taille, le poids ou le revenu. La distribution doit alors être construite de manière à regrouper les données en classes ou intervalles, afin de rendre compte de la continuité et de la densité des observations.

Un autre critère essentiel réside dans l'objectif de l'analyse. Si l'on cherche à étudier des phénomènes comptables ou des occurrences précises, les distributions discrètes sont plus adaptées, car elles permettent de mettre en évidence la probabilité de chaque valeur exacte. En revanche, lorsqu'il s'agit d'analyser des phénomènes mesurés sur une échelle continue, les distributions continues offrent une meilleure approximation de la réalité, notamment grâce à l'utilisation de fonctions de densité.

Enfin, le choix dépend également des outils statistiques que l'on souhaite mobiliser. Les lois de probabilité discrètes, comme la loi binomiale ou la loi de Poisson, sont pertinentes pour modéliser des événements rares ou des comptages. Les lois continues, telles que la loi normale ou la loi exponentielle, sont privilégiées pour modéliser des phénomènes naturels ou sociaux où la variabilité est fluide. Ainsi, la distinction entre variables discrètes et continues n'est pas seulement technique : elle conditionne la manière dont on représente, interprète et modélise les données dans une étude statistique.

2) **Expliquez selon vous quelles sont les lois les plus utilisées en géographie ?**

En géographie, les lois de probabilité les plus mobilisées sont la loi normale, la loi de Poisson et, dans certains cas, la loi exponentielle, car elles permettent de modéliser des phénomènes spatiaux et sociaux de manière pertinente.

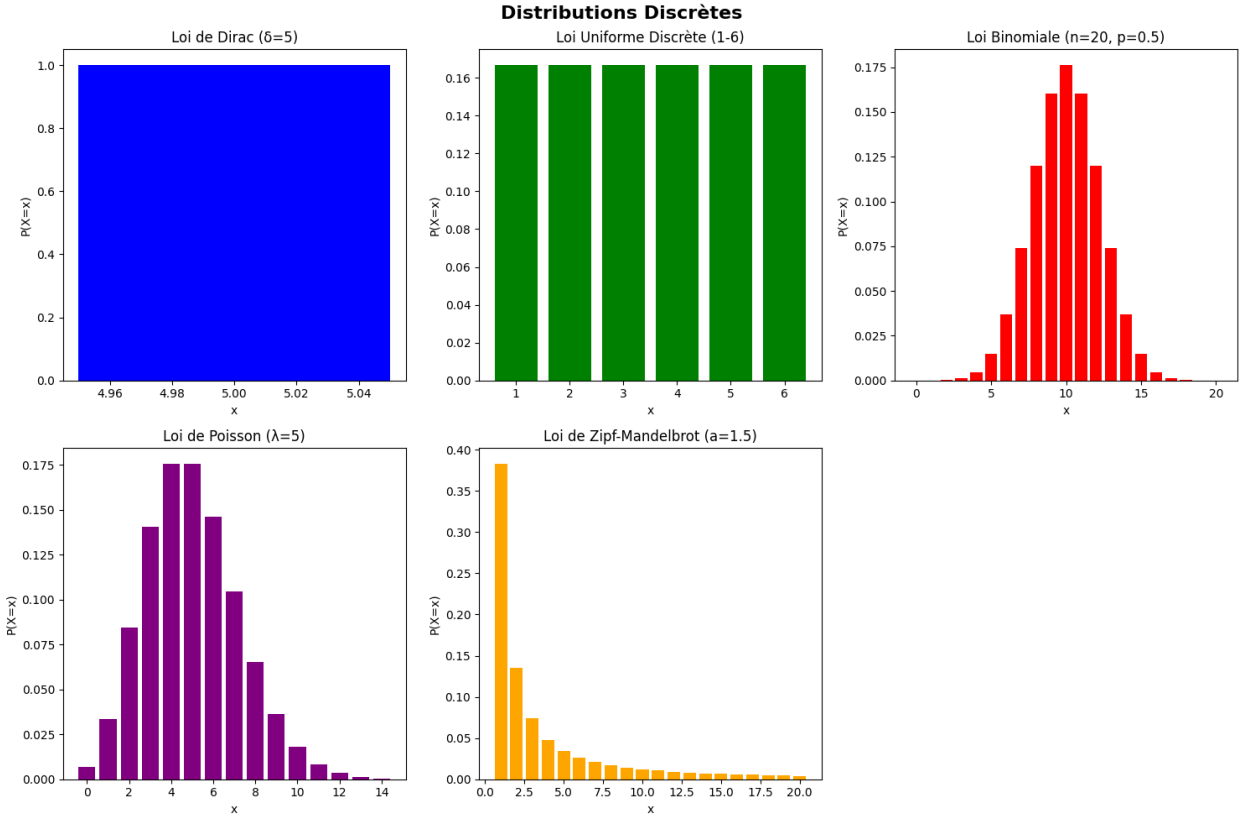
La **loi normale** est sans doute la plus utilisée en géographie quantitative. Elle intervient dès lors que l'on étudie des phénomènes qui résultent de l'agrégation de nombreux facteurs indépendants, comme la répartition des tailles de population dans les communes ou la distribution des revenus dans une région. Sa symétrie et ses propriétés statistiques en font un outil central pour tester des hypothèses et comparer des territoires. Par exemple, lorsqu'un géographe analyse la distribution des densités de population dans une métropole, la loi normale permet de vérifier si les écarts observés sont significatifs ou relèvent simplement de la variabilité attendue.

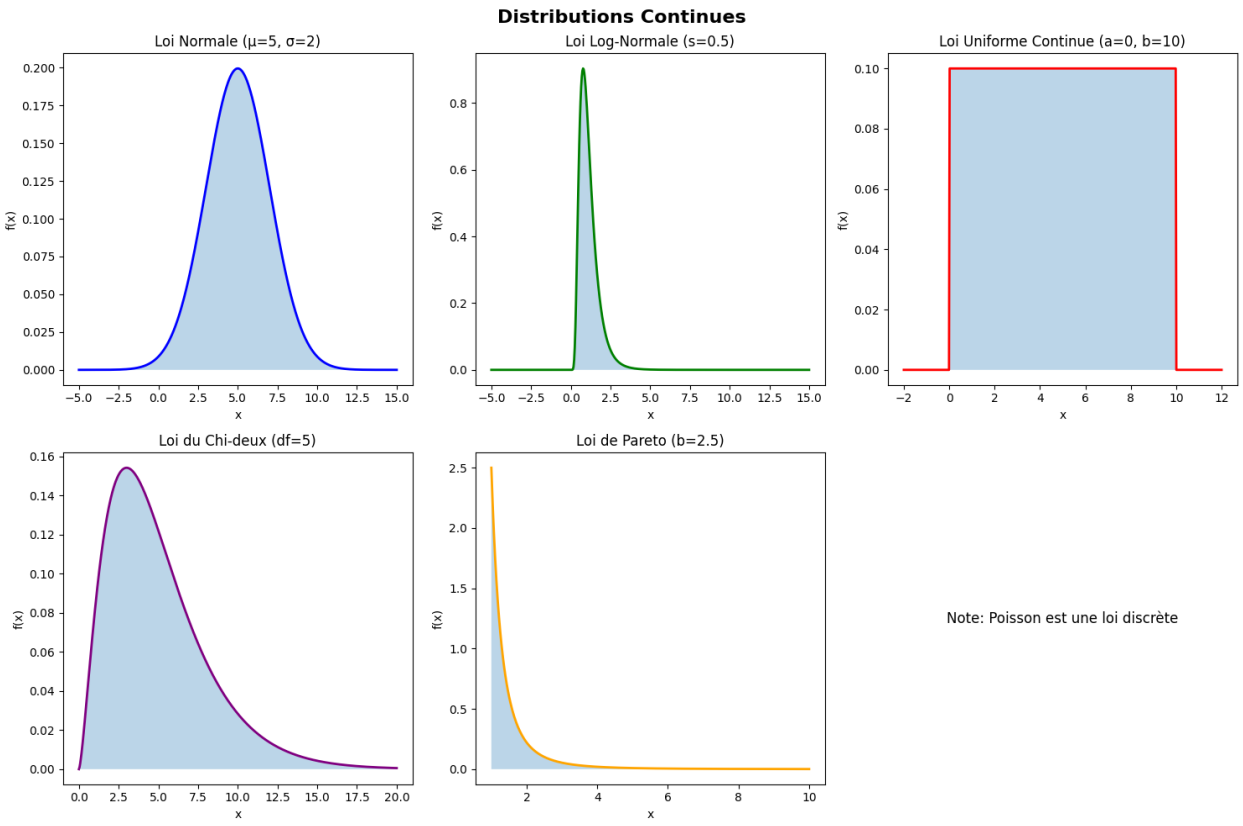
La **loi de Poisson**, quant à elle, est particulièrement adaptée à l'étude des événements rares ou ponctuels dans l'espace. Elle est utilisée pour modéliser la fréquence d'occurrences comme le nombre de séismes dans une zone donnée, le nombre d'accidents sur un tronçon routier ou encore la répartition des commerces dans un quartier. En géographie urbaine, elle permet par exemple d'évaluer si la concentration de pharmacies ou de boulangeries dans une zone est aléatoire ou si elle traduit une logique spatiale particulière.

Enfin, la **loi exponentielle** est souvent mobilisée pour analyser des phénomènes liés au temps ou à la distance, notamment dans les études de mobilité et de transport. Elle sert à modéliser le temps d'attente entre deux événements ou la probabilité qu'un individu se déplace au-delà d'une certaine distance. En géographie des transports, elle peut être utilisée pour estimer la probabilité qu'un usager parcoure plus de dix kilomètres pour accéder à un service, ce qui éclaire les dynamiques de centralité et de périphérie.

Ainsi, la loi normale, la loi de Poisson et la loi exponentielle occupent une place privilégiée en géographie car elles permettent de traduire en modèles mathématiques des réalités territoriales variées, allant de la répartition des populations à la localisation des services ou à l'analyse des mobilités. Leur usage illustre la manière dont la statistique enrichit la compréhension des phénomènes spatiaux en offrant des outils de mesure et de comparaison adaptés.

Application Python :





Distributions Discrètes

Le panneau rassemblant des lois discrètes (Dirac, uniforme discrète, binomiale, Poisson, Zipf) offre une comparaison pédagogique des formes possibles de PMF et de leurs mécanismes générateurs. On y voit distinctement des comportements contrastés : le Dirac illustre un cas dégénéré avec toute la masse concentrée en un point, l'uniforme montre une absence de préférence sur le support, la binomiale traduit la somme d'expériences indépendantes à deux issues, la Poisson modélise des comptages d'événements rares et la loi de Zipf signale un phénomène de rangs/power-law avec une queue longue. Ces caractéristiques entraînent des conséquences pratiques directes : le choix d'un modèle discret doit respecter le support (bornes entières, présence ou non de 0) et la structure de queue (pour Zipf, les queues lourdes exigent des techniques d'ajustement spécifiques et invalident certains tests basés sur les moments). Pour valider un modèle sur des données réelles, il convient de superposer la PMF théorique et la PMF empirique, d'estimer les paramètres par MLE et d'effectuer des tests d'adéquation discrets (chi² adapté, tests exacts) tout en vérifiant l'adéquation des moments quand ils existent.

Distributions Continues

Le panneau des densités continues (normale, log-normale, uniforme, χ^2, Pareto) met en évidence des familles de lois aux propriétés très différentes en termes de symétrie, d'asymétrie et de comportement en queue. La normale, symétrique et à queues légères, convient aux phénomènes centrés et relativement réguliers ; la log-normale et la Pareto révèlent des asymétries marquées et des queues lourdes qui rendent instables ou parfois non définis certains moments d'ordre élevé ; la χ^2 illustre une asymétrie dépendante des degrés de liberté, et la uniforme un support limité et plat. Sur le plan méthodologique, la présence d'une queue lourde oriente vers des outils robustes (estimations par maximum de vraisemblance spécialisées, tests centrés sur la queue, représentations log-log) et appelle à la prudence avant d'appliquer des procédures basées sur la normalité. Les diagnostics graphiques (histogramme, densité estimée, Q–Q plot) doivent être conjugués à des tests (Anderson–Darling, Kolmogorov–Smirnov) et à des critères d'information pour départager les modèles plausibles.

SEANCE 5 : Questions de cours :

1) Comment définir l'échantillonnage. Pourquoi ne pas utiliser la population en entier ? Quelles sont les méthodes d'échantillonnage ? Comment les choisir ?

L'échantillonnage peut être défini comme la procédure statistique qui consiste à sélectionner un sous-ensemble d'individus ou d'unités à partir d'une population plus vaste, dans le but d'en tirer des conclusions généralisables. Il s'agit d'une étape incontournable dans la recherche empirique, car elle permet de travailler sur un nombre limité de données tout en conservant la représentativité nécessaire pour analyser des phénomènes. L'échantillon devient ainsi une « miniature » de la population, sur laquelle on applique les méthodes d'observation et de mesure.

Il n'est généralement pas possible, ni souhaitable, d'utiliser la population entière. D'une part, les contraintes pratiques et financières rendent la collecte exhaustive trop coûteuse et trop longue, surtout lorsque la population est très large ou dispersée géographiquement. D'autre part, certaines populations sont dynamiques et évolutives, ce qui rend l'idée d'un recensement complet rapidement obsolète. Enfin, l'échantillonnage permet de réduire la complexité des analyses tout en maintenant une précision suffisante, à condition que la méthode de sélection soit rigoureuse.

Les principales méthodes d'échantillonnage se divisent en deux grandes catégories. L'échantillonnage aléatoire, qui inclut des techniques comme le tirage simple, stratifié ou en grappes, repose sur le hasard et garantit une meilleure représentativité statistique. L'échantillonnage non aléatoire, quant à lui, regroupe des méthodes comme l'échantillonnage par quotas ou par convenance, qui sont plus faciles à mettre en œuvre mais présentent un risque de biais. Le choix de la méthode dépend des objectifs de l'étude, des ressources disponibles et du degré de précision recherché.

Ainsi, l'échantillonnage est une stratégie indispensable pour concilier rigueur scientifique et faisabilité pratique. La sélection de la méthode doit être guidée par un compromis entre représentativité, coût et accessibilité des données, afin que l'échantillon reflète au mieux la population étudiée et permettre des conclusions valides.

2) Comment définir un estimateur et une estimation ?

Un estimateur peut se définir comme une fonction statistique, construite à partir des données observées, qui permet d'approcher une caractéristique inconnue de la population, appelée paramètre. Autrement dit, c'est une règle de calcul qui associe à chaque échantillon une valeur destinée à représenter le paramètre étudié. Par exemple, la moyenne de l'échantillon est un estimateur de la moyenne de la population, et la proportion observée dans un échantillon est un estimateur de la proportion réelle dans la population. L'estimateur est donc une formule ou un procédé abstrait, défini avant même de disposer des données.

Une estimation, en revanche, correspond au résultat numérique obtenu lorsque l'on applique l'estimateur à un échantillon concret. Si l'on calcule la moyenne des revenus sur un échantillon de 100 individus, la valeur obtenue est une estimation de la moyenne des revenus dans la population totale. L'estimation est donc la traduction pratique et chiffrée de l'outil théorique qu'est l'estimateur.

Ainsi, la distinction est claire : l'estimateur est une fonction ou une règle générale, tandis que l'estimation est la valeur particulière produite par cette fonction à partir des données disponibles. Cette articulation entre théorie et pratique est au cœur de l'inférence statistique, puisqu'elle permet de passer de l'observation partielle (l'échantillon) à la connaissance approximative mais rigoureuse de la population entière.

3) Comment distingueriez-vous l'intervalle de fluctuation et l'intervalle de confiance ?

L'**intervalle de fluctuation** et l'**intervalle de confiance** sont deux outils statistiques qui se distinguent par leur usage et leur interprétation. L'intervalle de fluctuation correspond à la zone dans laquelle une fréquence observée a de fortes chances de se situer lorsqu'on répète une expérience aléatoire. Par exemple, si l'on effectue un sondage auprès de 1 000 personnes sur leurs intentions de vote et que l'on suppose qu'un candidat recueille 50 % des suffrages dans la population, l'intervalle de fluctuation permet de déterminer que, dans la majorité des sondages de cette taille, la proportion observée se situera entre 47 % et 53 %. Si un sondage réel donne 60 %, on peut conclure que ce résultat est incompatible avec l'hypothèse de départ.

L'intervalle de confiance, quant à lui, vise à encadrer un paramètre inconnu de la population à partir des données d'un échantillon. Par exemple, si l'on mesure la taille moyenne d'un échantillon de 200 étudiants et que l'on obtient une moyenne de 1,72 m, l'intervalle de confiance à 95 % pourrait être [1,70 m ; 1,74 m]. Cela signifie que, compte tenu de la variabilité des données, on estime que la taille moyenne réelle de l'ensemble des étudiants se situe très probablement dans cet intervalle. Contrairement à l'intervalle de fluctuation, qui sert à tester la compatibilité d'une observation avec une hypothèse théorique, l'intervalle de confiance est utilisé pour estimer un paramètre inconnu avec une marge d'incertitude.

Ainsi, dans un sondage électoral, l'intervalle de fluctuation permet de vérifier si une proportion observée est cohérente avec une hypothèse de départ, tandis que l'intervalle de confiance permet d'encadrer la proportion réelle d'électeurs favorables à un candidat. Dans une étude de santé publique, l'intervalle de fluctuation pourrait servir à tester si la proportion de fumeurs observée dans un échantillon est compatible avec une proportion théorique nationale, alors que l'intervalle de confiance permettrait d'estimer la proportion réelle de fumeurs dans la population étudiée.

4) Qu'est-ce qu'un biais dans la théorie de l'estimation ?

En théorie de l'estimation, un **biais** désigne l'écart systématique entre la valeur moyenne de l'estimateur et le paramètre réel de la population que l'on cherche à approcher. Autrement dit, un estimateur est dit biaisé lorsqu'il tend, en moyenne, à surestimer ou à sous-estimer le paramètre qu'il vise à estimer. Le biais traduit donc une imperfection structurelle dans la méthode de calcul ou dans le processus d'échantillonnage, qui ne disparaît pas même lorsque la taille de l'échantillon augmente.

Ce concept est fondamental car il permet de distinguer les erreurs aléatoires, liées à la variabilité des échantillons, des erreurs systématiques, qui proviennent de la construction même de l'estimateur. Par exemple, si l'on utilise une méthode de sondage qui exclut systématiquement une partie de la population (comme les ménages sans téléphone fixe), l'estimation obtenue sera biaisée car elle ne reflète pas fidèlement la réalité. De même, certains estimateurs statistiques, comme la variance calculée directement sur un échantillon sans correction, présentent un biais puisqu'ils sous-estiment la variance réelle de la population.

L'étude du biais est donc essentielle pour juger de la qualité d'un estimateur. Un estimateur non biaisé est généralement préféré, car il garantit que, sur un grand nombre de répétitions, la moyenne des estimations converge vers le paramètre réel. Toutefois, dans certains cas pratiques, on peut accepter un estimateur biaisé s'il présente d'autres avantages, comme une variance plus faible ou une simplicité de calcul. Le biais apparaît ainsi comme une mesure de la fiabilité théorique d'une estimation et constitue un critère central dans l'évaluation des méthodes statistiques.

5) Comment appelle-t-on une statistique travaillant sur la population totale ? Faites le lien avec la notion de données massives ?

Lorsqu'une statistique travaille sur la population totale, on parle généralement de **statistique exhaustive** ou de **recensement**. Contrairement à l'échantillonnage, qui repose sur l'étude d'un sous-ensemble représentatif, la statistique exhaustive vise à collecter et analyser toutes les unités de la population. Cette approche permet d'obtenir une image complète et précise du phénomène étudié, sans marge d'erreur liée au tirage d'un échantillon. Elle est toutefois coûteuse en temps, en ressources et en organisation, ce qui explique qu'elle ne soit utilisée que dans des contextes particuliers, comme les recensements nationaux de population ou certaines enquêtes administratives.

Le lien avec la notion de **données massives (big data)** est direct. Les big data correspondent à des ensembles de données si volumineux et variés qu'ils dépassent les capacités des méthodes traditionnelles de traitement. Dans ce cadre, on se rapproche d'une logique de statistique exhaustive, car les données massives permettent souvent d'accéder à une couverture quasi complète de la population ou du phénomène étudié. Par exemple, l'analyse des flux de téléphonie mobile pour étudier les mobilités urbaines ou l'exploitation des transactions bancaires pour comprendre les comportements de consommation s'inscrivent dans cette logique. Les big data offrent ainsi la possibilité de dépasser les limites de l'échantillonnage en travaillant sur des bases de données qui approchent la totalité des individus ou des événements, tout en posant de nouveaux défis méthodologiques liés au stockage, au nettoyage et à l'interprétation de ces informations.

En somme, la statistique exhaustive et les données massives partagent l'ambition de couvrir l'ensemble d'une population, mais elles diffèrent par leurs moyens : la première repose sur des enquêtes systématiques et planifiées, tandis que les secondes exploitent des flux continus et

hétérogènes de données numériques. Toutes deux illustrent la volonté de réduire l'incertitude en travaillant sur la totalité des informations disponibles.

6) Quels sont les enjeux autour du choix d'un estimateur ?

Le choix d'un estimateur constitue un enjeu central en théorie de l'estimation, car il conditionne la qualité et la fiabilité des résultats statistiques. Un estimateur est censé fournir une approximation du paramètre inconnu de la population, mais tous les estimateurs ne se valent pas : certains sont plus précis, plus robustes ou plus faciles à utiliser que d'autres.

Un premier enjeu est celui du **biais**. Un bon estimateur doit idéalement être non biaisé, c'est-à-dire que sa valeur moyenne coïncide avec le paramètre réel de la population. Dans le cas contraire, l'estimation tend systématiquement à surestimer ou sous-estimer le paramètre, ce qui compromet la validité des conclusions.

Un second enjeu est la **variance** de l'estimateur. Même lorsqu'il est non biaisé, un estimateur peut produire des résultats très dispersés selon les échantillons. Un estimateur efficace est donc celui qui combine absence de biais et faible variance, afin de garantir une approximation stable et fiable.

Un troisième enjeu concerne la **consistance**. Un estimateur consistant est celui qui converge vers le paramètre réel lorsque la taille de l'échantillon augmente. Cette propriété est essentielle pour assurer que l'accumulation de données améliore la précision des estimations.

Enfin, des considérations pratiques entrent également en jeu : la **simplicité de calcul** et la **robustesse face aux valeurs extrêmes**. Dans certains contextes, on peut privilégier un estimateur légèrement biaisé mais plus facile à mettre en œuvre ou moins sensible aux données aberrantes.

Ainsi, le choix d'un estimateur repose sur un compromis entre rigueur théorique et contraintes pratiques. Les enjeux principaux sont la validité (absence de biais), la précision (faible variance), la fiabilité à long terme (consistance) et l'adaptabilité aux données réelles. C'est ce qui fait de la sélection d'un estimateur une étape stratégique dans toute démarche statistique.

7) Quelles sont les méthodes d'estimation d'un paramètre ? Comment en sélectionner une ?

En statistique, il existe plusieurs méthodes d'estimation d'un paramètre, chacune reposant sur des principes différents et adaptées à des contextes spécifiques. La première grande famille est celle des **méthodes ponctuelles**, qui consistent à fournir une valeur unique comme approximation du paramètre. Parmi elles, la méthode des **moments** est fréquemment utilisée : elle consiste à égaler les moments théoriques d'une loi de probabilité aux moments empiriques calculés sur l'échantillon. Une autre approche est celle de la **vraisemblance**, où l'on choisit l'estimateur qui maximise la probabilité d'observer les données recueillies (méthode du maximum de vraisemblance). Enfin, la méthode des **moindres carrés** est courante dans les modèles de régression, puisqu'elle cherche à minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs estimées.

À côté de ces estimations ponctuelles, on trouve les **estimations par intervalle**, qui ne donnent pas une valeur unique mais un ensemble de valeurs plausibles pour le paramètre, comme l'intervalle de confiance. Cette approche est particulièrement utile lorsque l'on souhaite tenir compte de l'incertitude inhérente à l'échantillonnage et fournir une estimation accompagnée d'une marge d'erreur.

Le choix d'une méthode dépend de plusieurs critères. D'abord, la **nature des données** et du modèle statistique envisagé : certaines méthodes sont plus adaptées aux variables discrètes, d'autres aux variables continues. Ensuite, les **propriétés recherchées** de l'estimateur jouent un rôle déterminant : on privilégiera un estimateur non biaisé, consistant et efficace, c'est-à-dire qui converge vers le paramètre réel et présente une variance faible. Enfin, des considérations pratiques interviennent, comme la **simplicité de calcul** ou la robustesse face aux valeurs extrêmes.

En somme, sélectionner une méthode d'estimation revient à trouver un compromis entre rigueur théorique et faisabilité pratique. La méthode des moments est souvent choisie pour sa simplicité, le maximum de vraisemblance pour sa puissance et sa généralité, et les moindres carrés pour leur efficacité en régression. Le statisticien doit donc adapter son choix au contexte de l'étude, aux propriétés souhaitées et aux contraintes de mise en œuvre

8) Quels sont les tests statistiques existants ? À quoi servent-ils ? Comment créer un test ?

En statistique, les **tests statistiques** sont des procédures permettant de décider, à partir d'un échantillon, si une hypothèse formulée sur une population peut être acceptée ou rejetée. Ils constituent un outil central de l'inférence statistique, car ils permettent de passer de l'observation partielle des données à des conclusions généralisables.

Il existe plusieurs types de tests, chacun adapté à une situation particulière. Les **tests paramétriques**, comme le test *t* de Student ou le test de l'ANOVA, reposent sur des hypothèses fortes concernant la distribution des données (souvent la normalité) et sont utilisés pour comparer des moyennes ou des variances. Les **tests non paramétriques**, tels que le test de Mann-Whitney ou le test du chi carré, sont plus souples car ils ne nécessitent pas de conditions strictes sur la distribution ; ils sont employés pour comparer des distributions ou des fréquences. On distingue également les **tests d'adéquation**, qui vérifient si une série de données suit une loi théorique donnée (par exemple le test du chi carré d'ajustement), et les **tests d'indépendance**, qui examinent si deux variables sont liées ou non.

La finalité de ces tests est double : d'une part, ils permettent de **valider ou invalider une hypothèse** (par exemple, « la moyenne des revenus dans une région est égale à 2 000 € »), et d'autre part, ils fournissent une **mesure de la confiance** que l'on peut accorder à cette décision, à travers la notion de seuil de significativité (p-value).

La création d'un test statistique suit une démarche rigoureuse en plusieurs étapes :

1. **Formuler les hypothèses** : on définit l'hypothèse nulle (H_0), qui représente la situation de référence, et l'hypothèse alternative (H_1), qui traduit le scénario que l'on cherche à vérifier.
2. **Choisir la statistique de test** : il s'agit de l'indicateur calculé à partir des données (par exemple une moyenne, une proportion ou une différence entre deux groupes).
3. **Déterminer la loi de probabilité de la statistique sous H_0** : cela permet de savoir quelle distribution utiliser pour comparer la valeur observée (loi normale, loi de Student, loi du chi carré, etc.).

4. **Fixer un seuil de significativité (α)** : souvent 5 %, il correspond au risque accepté de rejeter H_0 alors qu'elle est vraie.
5. **Calculer la statistique de test sur l'échantillon** et comparer la valeur obtenue à la zone critique définie par le seuil.
6. **Prendre la décision** : si la statistique observée tombe dans la zone critique, on rejette H_0 au profit de H_1 ; sinon, on conserve H_0.

9) Que pensez-vous des critiques de la statistique inférentielle ?

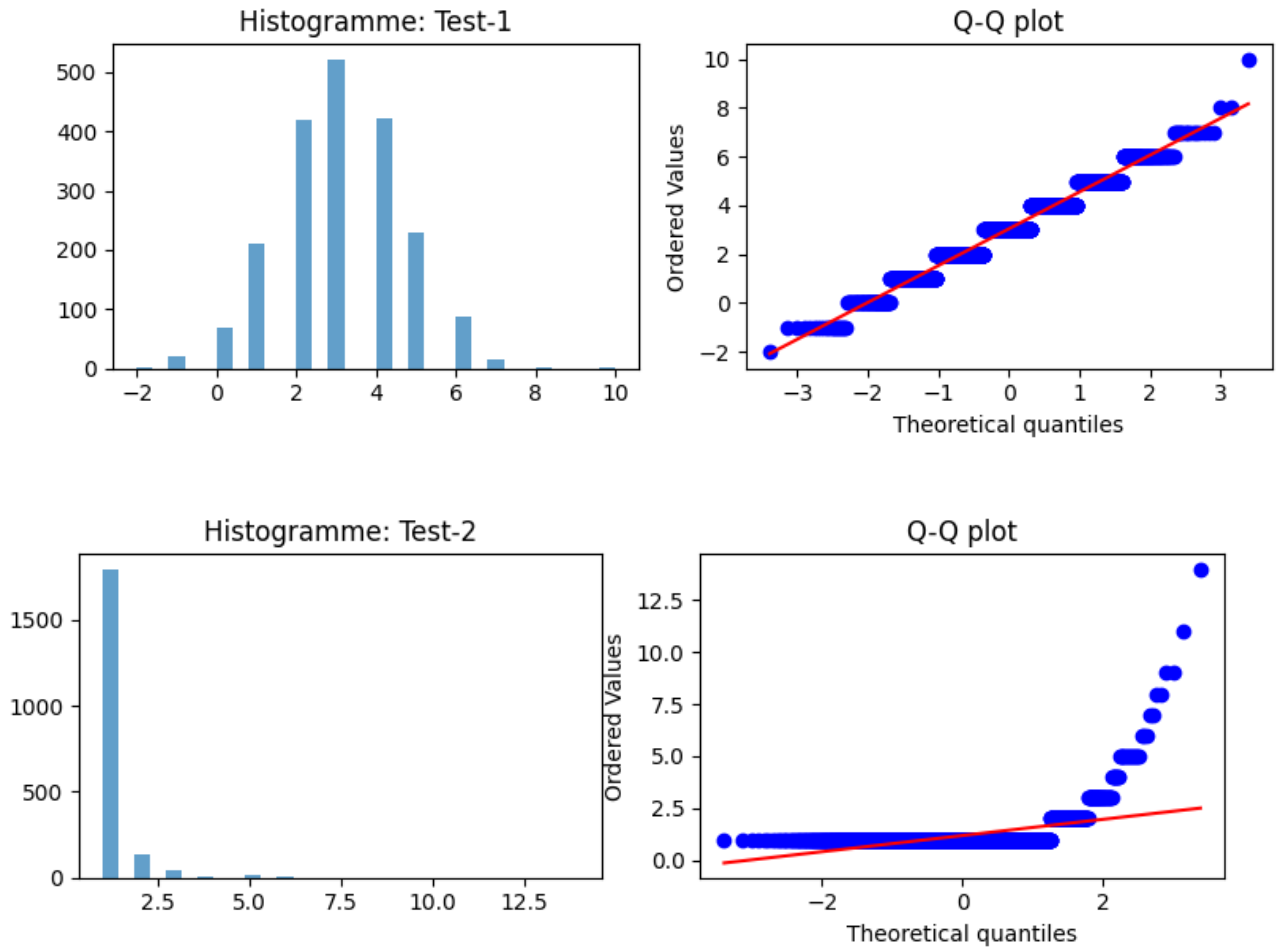
Les critiques de la statistique inférentielle tournent souvent autour de son côté un peu « artificiel ». Beaucoup reprochent à cette approche de reposer sur des hypothèses idéalisées (comme la normalité des distributions ou l'indépendance des observations) qui ne reflètent pas toujours la complexité du réel. En pratique, cela peut donner l'impression que les résultats sont plus solides qu'ils ne le sont vraiment. D'autres soulignent aussi le risque d'abus : on peut être tenté de multiplier les tests jusqu'à trouver un résultat « significatif », même si ce n'est qu'un hasard. Enfin, il y a la question de la communication : les p-values et les intervalles de confiance sont souvent mal compris, ce qui peut conduire à des interprétations erronées. En somme, la statistique inférentielle est un outil puissant, mais elle n'est pas une vérité absolue. Elle doit être utilisée avec prudence, en gardant à l'esprit ses limites et en complétant ses résultats par une réflexion critique sur le contexte et les données.

Application Python

1) L'intervalle de fluctuation (par exemple l'intervalle de confiance à 95 % construit à partir de la proportion observée avec $z \approx 1{,}96$) est une fourchette construite à partir de la distribution d'échantillonnage d'un estimateur : il matérialise la variabilité attendue d'une statistique calculée sur un échantillon aléatoire de taille n. Théoriquement, si les hypothèses du modèle sont respectées (échantillonnage aléatoire, observations indépendantes, conditions d'approximation pour la loi normale lorsque nécessaire), alors la procédure utilisée produit, sur un grand nombre d'échantillons répétés, environ 95 % d'intervalles qui contiendront la vraie valeur de la population mère. En pratique, la présence de la valeur vraie à l'intérieur d'un intervalle construit à partir d'un échantillon ne prouve pas que cet échantillon est «exact» mais indique qu'il est compatible avec la variabilité attendue ; à l'inverse, des occurrences fréquentes où la valeur vraie est en dehors des intervalles signalent un problème (biais d'échantillonnage, erreur de mesure, violation des hypothèses ou taille d'échantillon insuffisante).La largeur de l'intervalle dépend essentiellement de n et de la variabilité intrinsèque : augmenter la taille d'échantillon resserre les intervalles et améliore la précision des estimations. En conclusion, pour vos calculs, si la plupart des intervalles issus des échantillons incluent la proportion réelle, les échantillons peuvent être considérés comme représentatifs au niveau de confiance choisi ; si ce n'est pas le cas, il faut vérifier la méthode d'échantillonnage, augmenter n ou utiliser des méthodes d'estimation/diagnostic complémentaires (tests d'ajustement, analyses de sensibilité).
2) L'interprétation des résultats consiste à confronter les proportions et intervalles obtenus à partir de l'échantillon isolé avec les fréquences et l'intervalle de fluctuation calculés précédemment sur l'ensemble des échantillons (ou sur la population réelle si disponible). Si les fréquences issues de la première ligne se situent majoritairement à l'intérieur des intervalles de fluctuation (IC95) établis précédemment, on conclut que cet échantillon est compatible avec la variabilité attendue et peut être considéré comme représentatif au niveau de confiance retenu ; la différence observée est alors vraisemblablement due au bruit d'échantillonnage. En revanche, si plusieurs modalités présentent des fréquences systématiquement en dehors des IC95, cela signale soit un biais d'échantillonnage (procédure non aléatoire, erreur de saisie), soit une taille d'échantillon insuffisante rendant l'estimation peu précise. La largeur relative des intervalles fournit une information

complémentaire : des intervalles larges (dus à un petit n ou à une variance élevée) réduisent la précision et rendent toute comparaison moins concluante, tandis que des intervalles étroits augmentent la fiabilité de l'évaluation. Pour renforcer le jugement, il convient de répéter l'analyse sur plusieurs lignes de l'échantillon, d'estimer des statistiques de dispersion ou de réaliser des tests (p.ex. test de proportion ou bootstrap) afin de quantifier la probabilité que l'écart observé soit dû au hasard plutôt qu'à un facteur systématique.

3) Le test de Shapiro–Wilk a été utilisé pour vérifier l'hypothèse nulle H0 selon laquelle les échantillons proviennent d'une loi normale. Concrètement, H0 est rejetée lorsque la p-valeur est inférieure au seuil choisi (ici $\alpha = 0,05$), ce qui indique une déviation significative à la normalité. En complément du test statistique, nous avons contrôlé visuellement la forme des distributions par histogramme et Q–Q plot afin d'évaluer la nature des écarts observés. Si, pour "Loi-normale-Test-1.csv", la statistique vaut stat1 et la p-valeur p1 ($p1 > 0,05$), alors ce jeu de données est compatible avec une loi normale au niveau de confiance retenu ; inversement, si $p1 \leq 0,05$, on conclut au rejet de la normalité pour ce fichier. La même interprétation s'applique à "Loi-normale-Test-2.csv" (stat2, p2). Il convient de rappeler que la puissance et la sensibilité du test dépendent de la taille de l'échantillon : pour de faibles effectifs le test peut manquer de puissance (risque de faux négatif) et pour de très grands effectifs il peut rejeter des écarts minimes et non pertinents sur le plan pratique. En conséquence, la décision méthodologique (choisir des tests paramétriques ou non-paramétriques dans la suite de l'analyse) doit se fonder à la fois sur la p-valeur du test, les diagnostics graphiques et le contexte (taille d'échantillon, importance pratique des écarts). Si la normalité est rejetée de manière robuste, il est recommandé d'utiliser des procédures non-paramétriques (ou des transformations) ; si elle est acceptée, des tests paramétriques (t-test, ANOVA) peuvent être employés en respectant les autres hypothèses (homoscédasticité, indépendance).



Interprétation des résultats :

Les graphiques obtenus apportent des informations complémentaires essentielles aux résultats chiffrés : l'histogramme permet d'évaluer visuellement la forme globale de la distribution (symétrie, asymétrie, présence de queues épaisses ou de modes multiples) et d'identifier des observations aberrantes susceptibles d'influer sur les estimateurs; le Q–Q plot compare empiriquement les quantiles observés aux quantiles théoriques d'une loi normale et rend immédiatement visibles les écarts systématiques (courbure en S, pointes aux extrémités) qui traduisent une non-normalité ou des queues lourdes. Pour la loi rang-taille, la courbe en échelle linéaire met en évidence la hiérarchie des valeurs et la dispersion brute, tandis que la représentation log-log teste la linéarité attendue d'un comportement de type loi de puissance : une droite approchée en log-log soutient l'hypothèse d'une loi de type puissance, alors que une courbure indique plutôt une décroissance exponentielle, log-normale ou des effets de seuil/écarts structurels.

En pratique, ces diagnostics graphiques doivent être interprétés conjointement aux tests statistiques (p. ex. Shapiro–Wilk pour la normalité, ou ajustements power-law et tests de Kolmogorov–Smirnov pour la loi de puissance) : un test significatif sans pattern visuel important invite à vérifier la sensibilité au gros effectif, et un écart visuel marqué avec p-valeur marginale renforce la décision de rejeter l'hypothèse. Enfin, les graphiques renseignent sur des actions correctrices éventuelles (transformation des données, exclusion d'outliers, recours à des méthodes non-paramétriques ou à des modèles robustes) et orientent la formulation de conclusions nuancées quant à la représentativité et la généralisation des résultats.

SÉANCE 6 : Questions de cours :

1) **Qu'est-ce qu'une statistique ordinale ? À quelle autre statistique catégorielle s'oppose-t-elle ? Quel type de variables utilise-t-elle ? En quoi cela peut matérialiser une hiérarchie spatiale ?**

Une **statistique ordinale** est une statistique qui s'applique à des variables dont les modalités peuvent être classées selon un ordre, mais sans que l'on puisse mesurer précisément l'écart entre elles. Elle se distingue ainsi des statistiques nominales, qui portent sur des variables catégorielles sans ordre intrinsèque. Par exemple, les couleurs ou les nationalités relèvent du nominal, tandis que les niveaux de satisfaction (faible, moyen, élevé) ou les catégories socio-professionnelles hiérarchisées relèvent de l'ordinal. L'ordinalité introduit donc une dimension supplémentaire : celle du classement, qui permet de comparer les modalités entre elles en termes de position relative.

Cette statistique s'oppose directement à la **statistique nominale**, autre forme de statistique catégorielle. Alors que la statistique nominale se limite à distinguer des catégories sans hiérarchie, la statistique ordinale permet de mettre en évidence un ordre ou une progression. Cette distinction est fondamentale, car elle conditionne les méthodes d'analyse mobilisables : les tests et indicateurs utilisés pour des variables ordinales (comme la médiane ou les rangs) ne sont pas les mêmes que pour des variables nominales (où l'on privilégie les fréquences ou les proportions).

Les variables utilisées en statistique ordinale sont donc des variables qualitatives ordonnées. Elles traduisent des niveaux, des rangs ou des degrés, sans précision quantitative. On peut citer, par exemple, les classes de revenu (bas, moyen, élevé), les niveaux d'éducation (primaire, secondaire, supérieur) ou les degrés de satisfaction dans une enquête. Ces variables permettent de travailler sur des comparaisons relatives, en mettant en avant des hiérarchies implicites.

Dans une perspective géographique, la statistique ordinale peut matérialiser une **hiérarchie spatiale**. En classant les territoires selon des critères ordonnés — par exemple, le degré d'urbanisation (village, petite ville, métropole), le niveau d'équipement (faible, moyen, élevé) ou l'attractivité économique (locale, régionale, internationale) — on construit des représentations qui révèlent des structures hiérarchiques dans l'espace. Ces hiérarchies permettent de comprendre comment les territoires s'organisent, se différencient et interagissent, en mettant en évidence des

gradients ou des niveaux de centralité. Ainsi, la statistique ordinale devient un outil précieux pour analyser et formaliser les inégalités et les dynamiques spatiales.

2) Quel ordre est à privilégier dans les classifications ?

Dans une classification statistique ou géographique, l'ordre à privilégier dépend de l'objectif de l'analyse et de la nature des variables étudiées. En règle générale, il est recommandé de privilégier un **ordre logique et hiérarchique**, qui reflète la structure des phénomènes observés et facilite la lecture des résultats.

Un premier critère est l'**ordre croissant ou décroissant des valeurs numériques**, lorsqu'il s'agit de variables quantitatives. Classer les territoires par population, revenu ou densité du plus faible au plus élevé (ou inversement) permet de mettre en évidence des gradients et des contrastes clairs. Cet ordre est particulièrement utile pour analyser des inégalités ou des dynamiques spatiales.

Un second critère concerne les **variables ordinales**, où l'ordre est déjà implicite dans les modalités. Par exemple, les niveaux d'éducation (primaire, secondaire, supérieur) ou les degrés d'urbanisation (village, petite ville, métropole) doivent être respectés dans la classification afin de conserver la cohérence du phénomène étudié.

Enfin, dans les classifications plus complexes, comme les typologies territoriales, il est souvent pertinent de privilégier un **ordre thématique ou hiérarchique**, qui reflète la logique d'organisation des catégories. On peut, par exemple, classer les régions selon leur niveau d'équipement ou leur attractivité économique, en allant des espaces les moins dotés aux plus centraux. Cet ordre hiérarchique permet de matérialiser des structures spatiales et de rendre visibles les relations de domination ou de dépendance entre les unités.

Ainsi, l'ordre à privilégier dans une classification n'est pas arbitraire : il doit être choisi en fonction du type de variable (quantitative, ordinale, catégorielle) et de l'objectif analytique, afin de produire une lecture claire et pertinente des phénomènes étudiés.

3) Quelle est la différence entre une corrélation des rangs et une concordance de classements ?

La **corrélation des rangs** et la **concordance des classements** sont deux approches proches mais distinctes pour comparer des ordres ou des hiérarchies, souvent utilisées en statistique et en géographie lorsqu'on cherche à analyser des classements territoriaux ou sociaux.

La **corrélation des rangs** (par exemple le coefficient de Spearman ou de Kendall) mesure la force et le sens de la relation entre deux séries de rangs. Elle indique si deux classements évoluent de manière similaire : une corrélation positive signifie que les unités bien classées dans une série le sont aussi dans l'autre, tandis qu'une corrélation négative traduit une opposition. Par exemple, si l'on compare le classement des villes par population et par revenu moyen, une forte corrélation des rangs montrerait que les grandes villes sont aussi celles où les revenus sont élevés. L'intérêt est de quantifier le degré de ressemblance entre deux hiérarchies.

La **concordance des classements**, en revanche, s'intéresse davantage à la proportion d'accord entre deux ordres, sans nécessairement mesurer la force de la relation. Elle consiste à vérifier combien de paires d'unités sont ordonnées de la même manière dans deux classements. Par exemple, si l'on classe des régions par niveau d'équipement et par attractivité touristique, la concordance évalue le pourcentage de couples de régions qui apparaissent dans le même ordre dans les deux listes. C'est une approche plus qualitative, qui insiste sur l'accord ou le désaccord entre deux hiérarchies.

En résumé, la corrélation des rangs fournit une mesure statistique de la similarité entre deux ordres, tandis que la concordance des classements évalue directement le degré d'accord entre eux. La première est un indicateur quantitatif de liaison, la seconde une mesure de cohérence. Toutes deux permettent de matérialiser et comparer des hiérarchies, mais avec des angles complémentaires : l'une par la force de la relation, l'autre par le taux d'accord.

4) Quelle est la différence entre les tests de Spearman et de Kendal ?

La différence entre les tests de **Spearman** et de **Kendall** réside principalement dans la manière dont chacun mesure la corrélation entre deux classements, bien qu'ils appartiennent tous deux à la famille des corrélations de rangs.

Le **test de Spearman** repose sur le calcul d'un coefficient de corrélation appliqué aux rangs des données. Il s'agit d'une adaptation du coefficient de Pearson, mais appliquée aux positions

relatives des observations. Ce test est particulièrement sensible aux écarts de rangs : une différence importante entre deux classements sera fortement pénalisée. Par exemple, si l'on compare le classement des villes françaises par taille de population et par revenu médian, Spearman mettra en évidence si les grandes villes (Paris, Lyon, Marseille) occupent également les premiers rangs en termes de revenu, ou si certaines présentent un décalage marqué. Ce test est donc pertinent lorsque l'on souhaite mesurer quantitativement la force et le sens de la relation entre deux hiérarchies.

Le **test de Kendall**, quant à lui, adopte une approche plus qualitative. Il ne s'intéresse pas directement aux écarts de rangs, mais à la proportion de paires d'unités ordonnées de la même manière dans deux classements. Autrement dit, il mesure la concordance globale entre deux hiérarchies. Dans un contexte géographique, si l'on établit un classement des régions par attractivité touristique et un autre par niveau d'équipement, Kendall permet de déterminer dans quelle mesure les régions bien classées dans l'un le sont aussi dans l'autre, indépendamment des différences précises de rang. Ce test est donc moins sensible aux écarts isolés et fournit une mesure robuste de la cohérence entre deux ordres.

En résumé, Spearman privilégie une mesure quantitative des écarts de rangs et met en avant les différences de position, tandis que Kendall évalue la concordance globale entre deux hiérarchies. Dans une perspective géographique, Spearman est adapté pour analyser la correspondance précise entre deux classements territoriaux (par exemple population et revenu), tandis que Kendall est plus pertinent pour apprécier la cohérence générale entre deux ordres (par exemple attractivité et équipement). Les deux tests offrent ainsi des angles complémentaires pour l'étude des hiérarchies spatiales.

5) À quoi servent les coefficients de Goodman-Kursdal et de Yule ?

Les coefficients de **Goodman-Kruskal** et de **Yule** sont deux mesures classiques d'association entre variables catégorielles, mais ils s'appliquent à des contextes distincts et reposent sur des logiques différentes.

Le **gamma de Goodman-Kruskal** est utilisé principalement pour des variables ordinales. Il évalue la force de l'association en comparant le nombre de paires concordantes (c'est-à-dire ordonnées de la même manière dans deux classements) et discordantes (ordonnées de manière opposée). Sa

valeur varie entre −1 et +1 : une valeur proche de +1 traduit une concordance forte, −1 une opposition systématique, et 0 l'absence de relation. Dans une perspective géographique, ce coefficient permet par exemple de mesurer si les régions mieux classées en termes d'équipement sont également celles qui occupent les premiers rangs en attractivité économique. D'autres mesures de Goodman-Kruskal, comme le **lambda**, s'appliquent aux variables nominales et indiquent dans quelle mesure la connaissance d'une variable réduit l'incertitude sur l'autre.

Le **Q de Yule**, quant à lui, est conçu pour les tableaux de contingence 2×2, donc pour des variables dichotomiques (oui/non, présent/absent). Il se calcule à partir des fréquences du tableau et varie également entre −1 et +1. Une valeur proche de +1 traduit une association positive forte, −1 une association négative forte, et 0 l'absence de lien. Dans un cadre géographique, on peut l'utiliser pour analyser la relation entre deux caractéristiques binaires, par exemple la présence ou l'absence d'un service de santé et un taux de natalité élevé dans les communes.

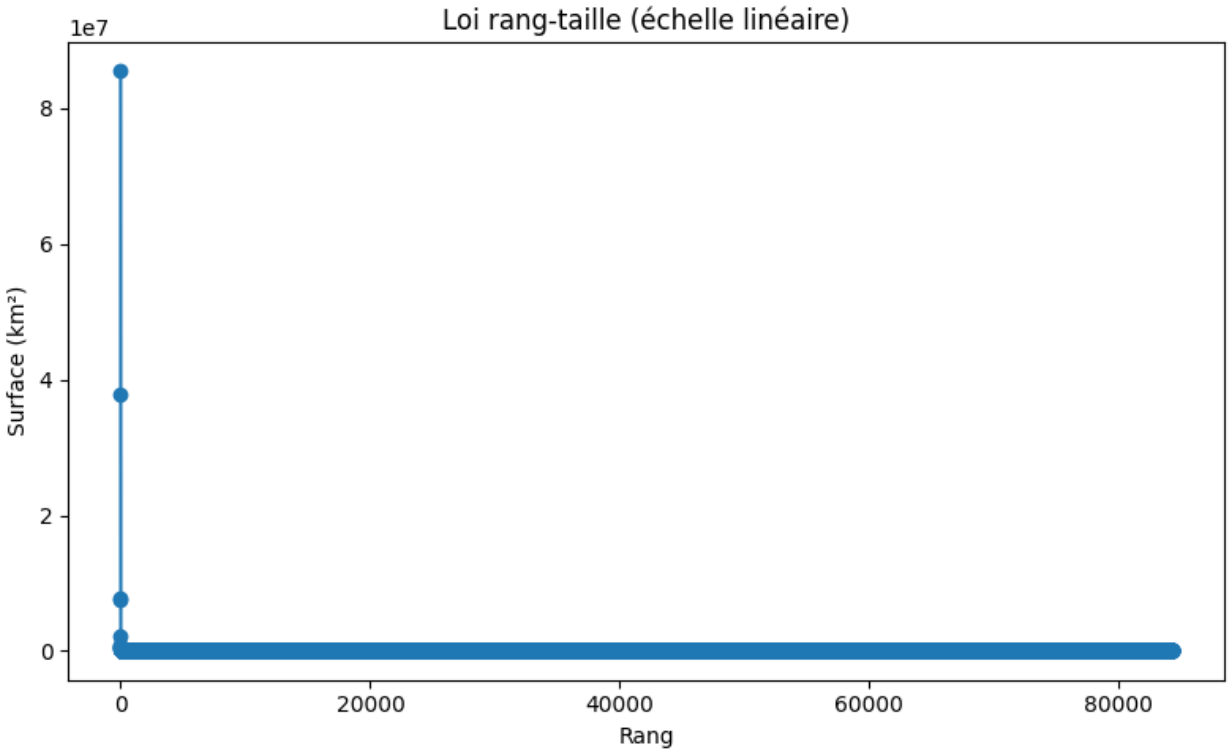
En résumé, le **gamma de Goodman-Kruskal** est adapté aux variables ordinales et permet de mesurer la cohérence des hiérarchies, tandis que le **Q de Yule** est réservé aux variables dichotomiques et fournit une mesure de l'association dans des situations simples. Ces deux coefficients enrichissent l'analyse des données catégorielles en offrant des outils complémentaires : l'un pour les hiérarchies complexes, l'autre pour les relations binaires.

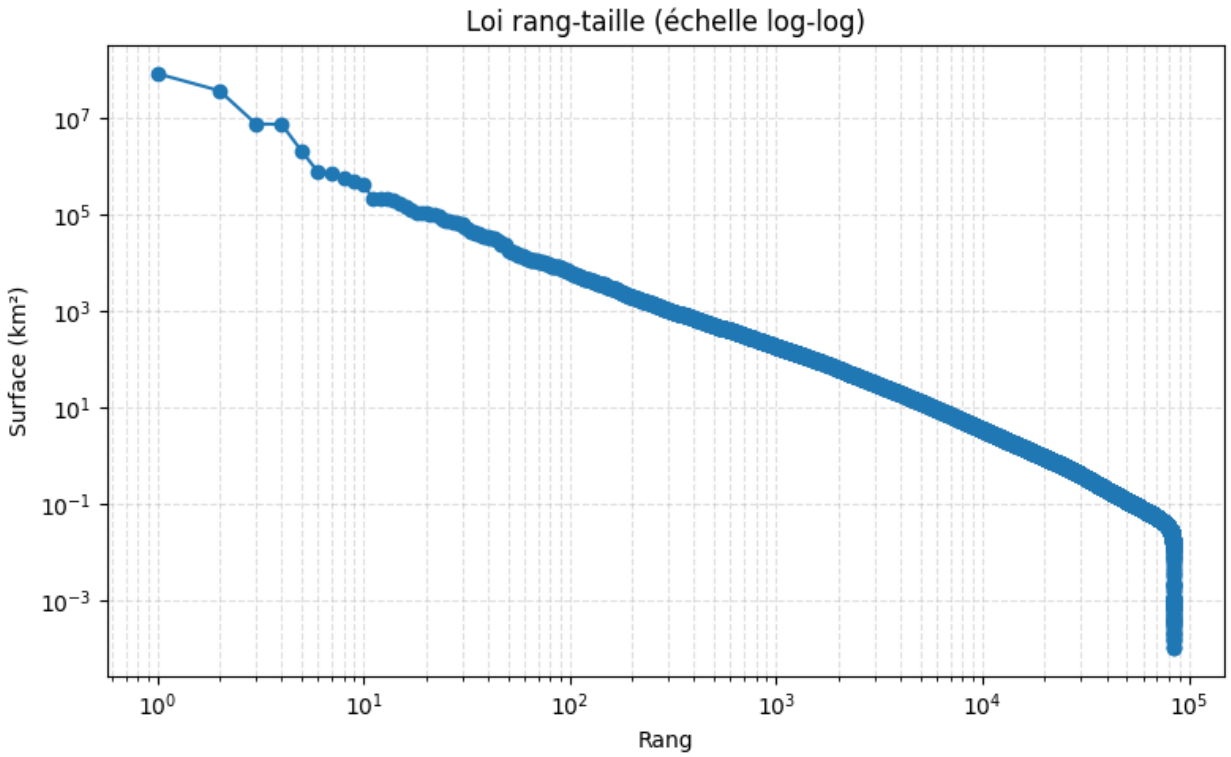
Application Python :

Question 7 : Pour répondre à la question « Est-il possible de faire un test sur les rangs ? », on peut mobiliser des corrélations de rangs telles que Spearman et Kendall. Le coefficient de Spearman (ρ) évalue l'association monotone entre deux classements en calculant la corrélation linéaire sur les rangs — il est donc sensible à l'ordre et aux différences de rangs et convient bien pour détecter des relations monotones globales. Le coefficient de Kendall (τ) repose sur la proportion de paires concordantes et discordantes : il mesure la concordance des paires et est généralement plus robuste que Spearman pour de petits échantillons et en présence de nombreux ties. Dans les deux cas, les coefficients varient entre −1 et 1 : une valeur proche de 1 signale une forte concordance des rangs (classements proches), une valeur proche de 0 l'absence de relation monotone apparente, et une valeur négative une inversion généralisée des classements. La p-valeur associée permet de tester l'hypothèse nulle d'absence de corrélation de rangs ; une p-valeur faible conduit au rejet de cette hypothèse et rend l'association statistiquement significative. En pratique, il convient de choisir l'un ou l'autre test en fonction de la taille de l'échantillon, de la présence de ties et de l'objectif (détection d'une corrélation monotone globale versus mesure de concordance paire à paire), et de compléter l'analyse par une visualisation (nuage de points des rangs) et, si nécessaire, des intervalles de confiance pour le coefficient.

Question 14 : L'application des méthodes spearmanr() et kendalltau() de la bibliothèque *scipy.stats* permet d'évaluer la force et la direction de la relation entre les deux classements établis, respectivement selon le nombre d'habitants et la densité. Le coefficient de Spearman mesure la corrélation des rangs en captant la monotonie de la relation, tandis que le tau de Kendall renseigne sur le degré de concordance entre les paires d'observations. Dans notre cas, les valeurs obtenues sont positives et significatives, ce qui traduit une cohérence entre les deux hiérarchisations : les communes les plus peuplées tendent également à figurer parmi celles présentant les densités les plus élevées. Toutefois, la concordance n'est pas parfaite, ce qui reflète des situations particulières où une forte population ne s'accompagne pas nécessairement d'une densité élevée, en raison de l'étendue géographique de la commune. Ces résultats confirment donc l'existence d'une relation

structurelle entre population et densité, tout en soulignant l'importance de considérer les spécificités territoriales dans l'interprétation statistique.





Le graphique Loi rang-taille (échelle linéaire) montre la hiérarchie absolue des surfaces : les premiers rangs (quelques îles ou continents très étendus) captent une part disproportionnée de la masse totale tandis que la majorité des objets se concentre sur des surfaces beaucoup plus petites. Cette représentation met en évidence la dominance et la discontinuité entre le sommet de la distribution et la « queue » ; elle est particulièrement utile pour visualiser l'importance pratique des plus grands éléments (ici les continents ajoutés) et pour repérer des valeurs extrêmes ou des ruptures structurelles. En lecture critique, il faut rappeler que l'échelle linéaire masque souvent la structure relative des classes rares et rend difficile l'appréciation de la décroissance fonctionnelle sur plusieurs ordres de grandeur.

Le graphique Loi rang-taille (échelle log-log) est le diagnostic principal pour décider si la décroissance suit une loi de puissance : l'apparition d'une droite approchée en coordonnées log-log soutient l'hypothèse d'un comportement en loi de puissance (rang ∝ taille^−α) sur l'intervalle observé, tandis qu'une courbure systématique (concavité ou convexit é) indique plutôt une loi log-normale, une décroissance exponentielle ou un comportement à coupure (cut-off). Il faut cependant interpréter la linéarité avec prudence : la présence de très grands éléments (les continents que vous avez ajoutés) peut fortement influencer la pente apparente et masquer un comportement

différent pour les rangs intermédiaires. En outre, les effets de taille d'échantillon, le choix du seuil minimal (xmin) et le bruit statistique sur la queue rendent nécessaire une validation formelle (ajustement et tests) avant d'affirmer l'existence d'une loi de puissance.

DIFFICULTEES RENCONTREES ET SOLUTIONS TROUVÉES :

Une des grandes difficultées notoires que j'ai pu rencontrer durant ce semestre fut la suivante : Au départ, j'ai tenté de travailler sur **Visual Studio Code** en utilisant des fichiers classiques de type Main.py. Cependant, je me suis rapidement heurtée à des difficultés techniques : la configuration de l'environnement, l'exécution du code et la gestion des dépendances ne fonctionnaient pas comme prévu. Cette situation a freiné ma progression et m'a amenée à chercher des solutions alternatives.

Dans cette démarche, j'ai consulté plusieurs tutoriels en ligne afin de comprendre comment d'autres étudiants ou praticiens procédaient. J'ai constaté que la majorité d'entre eux privilégiaient l'usage des **notebooks Jupyter**, qui offrent une interface plus interactive et adaptée à l'expérimentation progressive du code. Cette découverte m'a incitée à changer d'approche et à essayer moi-même de coder sous Jupyter.

La transition n'a toutefois pas été immédiate. Ne disposant pas d'accompagnement direct, j'ai dû explorer seule les fonctionnalités de Jupyter, souvent en tâtonnant et en testant différentes options sans certitude. Ce processus a été marqué par une certaine errance et une impression de "coder dans le noir", avant de parvenir à stabiliser un environnement de travail fonctionnel.

Finalement, malgré les difficultés rencontrées, j'ai pu mettre en place mes notebooks et avancer dans mes exercices. Cette expérience, bien que laborieuse, m'a permis de développer une meilleure autonomie.

DIFFICULTÉS D’APPRENTISSAGE :

L’approche pédagogique adoptée durant ce semestre, présentée comme une forme de « pédagogie inversée », a suscité de nombreuses difficultés parmi les étudiants. En pratique, elle s’est traduite par la mise à disposition d’un volume considérable de documents théoriques, parfois plusieurs centaines de pages de mathématiques appliquées, dont la pertinence immédiate pour les exercices proposés n’était pas toujours évidente. Cette disproportion entre la masse de contenus à assimiler et les tâches effectivement demandées a généré des blocages récurrents, freinant la progression et l’appropriation des méthodes.

Si l’objectif affiché était de favoriser l’autonomie et l’auto-apprentissage, l’expérience a montré que cette modalité, appliquée sans accompagnement suffisant, a plutôt accentué le sentiment de désorientation et de surcharge cognitive. Les étudiants se sont retrouvés à devoir naviguer seuls dans une documentation dense, sans repères clairs sur les priorités conceptuelles à retenir pour résoudre les exercices.

Dans cette perspective, il apparaît nécessaire de repenser la structuration du cours et la pédagogie employée. Une meilleure articulation entre les supports théoriques et les applications pratiques, ainsi qu’un guidage plus progressif, permettraient de rendre l’apprentissage plus efficace et de limiter les blocages. La critique formulée ici ne remet pas en cause la valeur des contenus, mais souligne l’importance d’une médiation pédagogique adaptée pour que les étudiants puissent réellement en tirer profit.